5-14-2024

Data-driven Evaluation of Deep Generative Models in Biomedical Imaging

Rucha Milind Deshpande
Washington University – McKelvey School of Engineering

WASHINGTON UNIVERSITY IN ST. LOUIS

McKelvey School of Engineering
Department of Biomedical Engineering

Dissertation Examination Committee:
Baranidharan Raman, Chair
Mark A. Anastasio, Co-Chair
Adam Bauer
Frank J. Brooks
Abhinav Jha
Quing Zhu

Data-driven Evaluation of Deep Generative Models in Biomedical Imaging

by

Rucha Milind Deshpande

A dissertation presented to
the McKelvey School of Engineering
of Washington University in
partial fulfillment of the
requirements for the degree
of Doctor of Philosophy

May 2024
St. Louis, Missouri

Table of Contents

List of Figures

List of Tables

List of Abbreviations

A-SCM: Alphabet - Stochastic Context Model

CNN: Convolutional Neural Network

DDPM: Denoising Diffusion Probabilistic Model

DGM: Deep Generative Model

DNN: Deep Neural Network

F-SCM: Flags - Stochastic Context Model

FID: Fréchet Inception Distance (A popular image quality metric for DGM images)

GAN: Generative Adversarial Network

LDM: Latent Diffusion Model

PG: ProGAN (An example of a generative adversarial network)

SCM: Stochastic Context Model

SG2: StyleGAN2 (An example of a generative adversarial network)

SOM: Stochastic Object Model

V-SCM: Voronoi - Stochastic Context Model

VICTRE: Virtual Imaging Clinical Trial for Regulatory Evaluation (An established virtual breast phantom)

VT-SOM: VICTRE - Stochastic Object Model (An adapted version of the established VICTRE phantom)

Acknowledgments

I am grateful to my advisor, Prof. Mark Anastasio, for broadening my scientific perspective with an understanding of imaging systems and science, for teaching me the importance of balancing emerging and conventional methods in science, and also for lessons in scientific leadership.

I am indebted to my co-advisor, Prof. Frank Brooks, for teaching me the importance of asking the right questions, for lessons in doing science right, and for inculcating many of my research habits. I am also grateful to him for enabling me to find joy and inspiration at the intersection of science, nature, and art. Most of all, I am grateful for his wisdom, kindness, and trust during tough times.

I would like to thank all members of my thesis committee for their feedback on my work, and Prof. Raman also for his support as my WashU mentor. I would especially like to thank Prof. Jha for introducing me to the fundamentals of imaging science, and for his professional advice over the last few years.

I am thankful to all my current and past lab-mates in the Anastasio lab, especially those with whom I have worked on projects together: Ashish Avachat, Varun Kelkar, and Muzaffer Özbey. I would particularly like to thank Seonyeong Park for her kindness within and outside the lab. I am also thankful to my collaborators at the U.S. FDA for sharing their diverse scientific perspectives.

I am indebted to all my mentors for expanding my horizons, sharing their wisdom, and giving honest feedback. I am especially grateful to Prof. Nachi Chockalingam for introducing me to the research lifestyle, and for his practical advice and encouragement over many years. I am also grateful to Cdr. (Retd.) M. R. Joglekar for his timely advice and support.

I am grateful to my friends for their joyful company, kindness, and honesty.

Finally, I thank my parents for being my strength, and my sister for her love.

I acknowledge all the funding sources that supported this work: NIH awards EB031585, EB034249, EB031772 (subproject 6366), CA238191, EB020604, EB023045, and the American Association of Physicists in Medicine (AAPM). This work utilized resources supported by the NSF grants 1725729, OCI 2005572, NCSA, UIUC, and the State of Illinois. I also acknowledge support from the WU-ISP Trainee Fellowship (5T32 EB01485505).

Rucha Milind Deshpande

Washington University in St. Louis
May 2024

To my parents, and to my mentors.

ABSTRACT OF THE DISSERTATION

Data-driven Evaluation of Deep Generative Models in Biomedical Imaging

by

Rucha Milind Deshpande

Doctor of Philosophy in Biomedical Engineering

Washington University in St. Louis, 2024

Professor Baranidharan Raman, Chair

Professor Mark A. Anastasio, Co-Chair

Deep generative models (DGMs) have tremendous potential for several biomedical imaging applications such as data augmentation, image reconstruction, and image denoising. However, the deployment of DGMs in real-world biomedical imaging workflows without domain-relevant evaluations can jeopardize patient health and well-being.

The evaluation of DGMs in biomedical imaging is challenging due to several factors: requirement of domain expertise for visual inspection, lack of a mathematically defined ground truth, and the unclear relevance of popular evaluation measures adopted from the computer vision literature. Given these challenges, one way to evaluate DGMs is via purposefully designed synthetic data. In this thesis, two frameworks for the evaluation of DGMs are proposed based on the idea of assessing reproducible "spatial context". Context is defined as domain-specific external knowledge that manifests as conditional co-occurrences of specific pixel arrangements in an image.

In the first of two frameworks, stochastic context models were purposefully designed to encode and assess the reproducibility of explicitly prescribed spatial context. Context was encoded in these models as contextual attributes such as per-image feature prevalence, feature-specific intensity distribution, and prescribed texture. In the second evaluation framework, a

more complex dataset: a stochastic model of the human female breast was adapted to evaluate DGMs for reproducible spatial context that arises implicitly due to structural variations in anatomy. All designed datasets are made publicly available to aid the benchmarking of novel and emerging DGMs.

The designed evaluation frameworks were employed to assess diffusion models, which are state-of-the-art DGMs and have been claimed to substantially outperform the other major DGM family: generative adversarial networks, in terms of visual image quality and popular evaluation measures. It was found that diffusion models hold promise for data augmentation tasks but errors may occur in the generation of multiple contextual attributes, and that popular evaluation measures do not capture these contextual errors.

From all studies, it was found that no modern DGM perfectly reproduced the expected spatial context. This highlights the need for further development of domain-specific DGMs as well as domain-relevant evaluation methods to ensure the safe and beneficial translation of DGM-based methods to real-world workflows in biomedical imaging.

Chapter 1

Introduction

"Science in the service of humanity is technology, but lack of wisdom may make the service harmful." - Isaac Asimov

A deep generative models (DGM) of images is a kind of a deep neural network (DNN) that can learn to synthesize images. DGMs have have been increasingly popular in medical imaging research for applications [1–4] such as dataset augmentation [5, 6], image denoising and superresolution [7], e.g., low-dose computed tomography (CT) to high-dose CT [8, 9], transforming images from one imaging modality into another [10–12], e.g., virtual histology staining [13], image reconstruction [14, 15], and image segmentation [16]. However, the capacity of DGMs to create images also enables the creation of unexpected artifacts, or "hallucinations", in the generated images. These hallucinations in DGM-generated biomedical images could substantially impact downstream decision support and potentially result in the loss of human lives. Therefore, it is imperative to evaluate DGMs before they are deployed in biomedical workflows. The evaluation of DGMs is challenging for many reasons and a major reason is that the generated images cannot be compared against a known "ground truth" in many scenarios. One approach to circumvent this challenge is via purposefully designed synthetic datasets relevant to biomedical imaging scenarios, thus creating a known ground truth. The present thesis focuses on developing synthetic-data-driven methods for the problem of DGM evaluation in biomedical imaging.

1.1 Scope of the Problem

A brief overview of the various applications of DGMs in biomedical imaging research is provided in the next subsection. This is followed by a discussion of hallucinations in DGM-generated images and the consequent potential risks, which highlight the need for evaluating DGMs with relevance to biomedical imaging.

1.1.1 Applications of DGMs in Biomedical Imaging Research

DGMs have found a wide range of applications in biomedical imaging research; some applications are listed below.

- Image synthesis for data augmentation: Often, biomedical image datasets are not sufficiently large to train learning-based methods (e.g., DNNs) for tasks such as classification and segmentation. Furthermore, certain pathologies might have very low prevalence in a dataset, and hence, may not be sufficiently represented for effective DNN training. DGM-generated ensembles can supplement the original biomedical dataset and thus, enable the training of DNNs (or even other automated methods) for downstream tasks. However, for such a strategy to be successful, it is essential that the diagnostic value of the training and DGM-generated ensembles be equivalent. DGMs have been employed for image synthesis applications such as the generation of brain magnetic resonance (MR) images [17, 18], chest radiographs [19, 20], and mammography images [21]. Sample DGM-generated images of brain MRI obtained from two DGMs trained on a popular brain MRI dataset [22] are shown in Figure 1.1. Note the high visual quality of the generated images and the lack of obvious artifacts to a non-domain-expert.

- Image denoising: Reduction of patient dose or image acquisition constraints may result in images with low signal-to-noise ratio (SNR). DGMs are being explored to transform these low quality images into high quality, denoised images. Examples include low-dose CT denoising [23] (see Figure 1.2 (a)), positron emission tomography (PET) denoising [24], and super-resolution of fundus retinal images [25].

Figure 1.1: Sample DGM-generated images of brain MRIs. Two DGMs trained on the fastMRI dataset [22] were employed to generate these images. No errors are obvious to a non-domain expert, and the images demonstrate high visual quality.

- Domain transfer: A DGM can learn to generate images from an imaging modality given images from another imaging modality; this is termed as domain transfer or more generally, image-to-image translation. Domain transfer could be useful when matched patient data is unavailable for an imaging modality, or when datasets from two imaging modalities are imbalanced and cannot be employed for downstream analysis or DNN training. Examples of domain transfer tasks include the generation of MR images from CT [26] (see Figure 1.2 (b)), and virtual staining of histopathology images [13].

- Image reconstruction: DGMs have been particularly popular for MR reconstruction [15, 27], but have also been employed for other applications such as limited-angle CT [28].

- Other applications: DGMs are being explored for many other applications including anomaly detection [29,30], image segmentation e.g., segmentation of blood vessels from images of the eye fundus [31] (see Figure 1.2 (c)), and image inpainting [32, 33], i.e., filling in missing regions in a given image. A wide overview of the various biomedical imaging applications of DGMs has been presented in some recent works [1–4].

1.1.2 Hallucinations in Biomedical Images

Typically, the use of any image involves the identification of certain task-relevant image features. An "image feature" maybe defined as a group or pixels or a derived measure

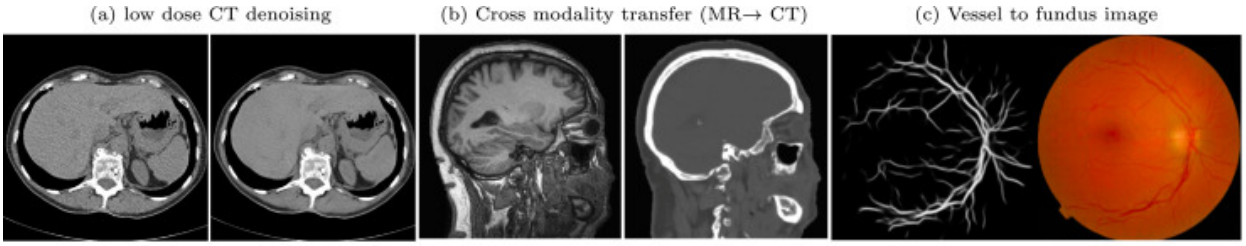


Figure 1.2: Examples of tasks performed by DGMs: (a) denoising of low-dose CT, (b) domain transfer or cross-modality generation (MR to CT), and (c) segmentation of blood vessels from images of the eye fundus. Reprinted from Yi et al. Copyright (2019) [1], with permission from Elsevier. Subfigures (a) [23] and (b) [26] reproduced with permission from Springer Nature, Copyright (2018, 2017), subfigure (c) [31] copyright (2017), IEEE.

identified in a certain manner. DGMs generate image features in a probabilistic manner to create new images. Although this probabilistic generation yields diversity and novelty, it can also lead to unexpected and unrealistic artifacts, i.e., hallucinations.

Muller et al. [34] have demonstrated instances of hallucinations in several biomedical imaging modalities as shown in Figure 1.3. For example, generated retinal images of the eye fundus demonstrated a gross anatomical error: two optical disks were present in the same image instead of the expected single disk present in all humans (Figure 1.3: row 1). A second kind of hallucination was observed as DGM-imposed texture in histology images, this unnatural texture can be identified even by a non-domain-expert (Figure 1.3: row 2). Last, in DGM-generated chest radiographs, major errors in the appearance and placement of support devices were observed, possibly because of the infrequent occurrence of these devices in the training images (Figure 1.3: row 3). If images with hallucinations are not identified and all generated images are employed for a downstream task, it is possible that a decision support system based on the generated ensemble might be inaccurate. The consequences of these inaccuracies might cascade via wrong diagnoses in the patient population, and potentially even cost patient lives.

Cohen et al. [35] also demonstrate that DGMs can cause massive hallucinations as a result of the composition of the target distribution and their training strategy. Specifically, when a DGM is trained via a distribution matching loss function, i.e., assessing whether the distributions of the training data and generated data are similar, the generated images are biased towards the target distribution. They demonstrate that in a domain transfer task, wherein MRI FLAIR images are transformed to T1 contrast images, a tumor may be added

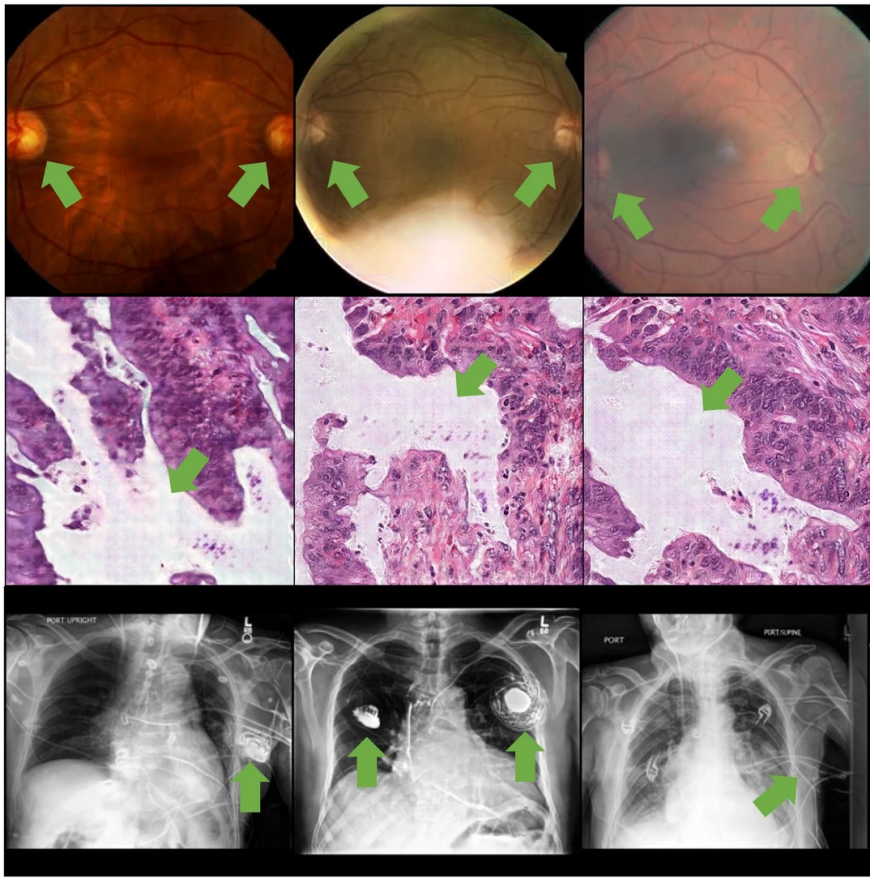
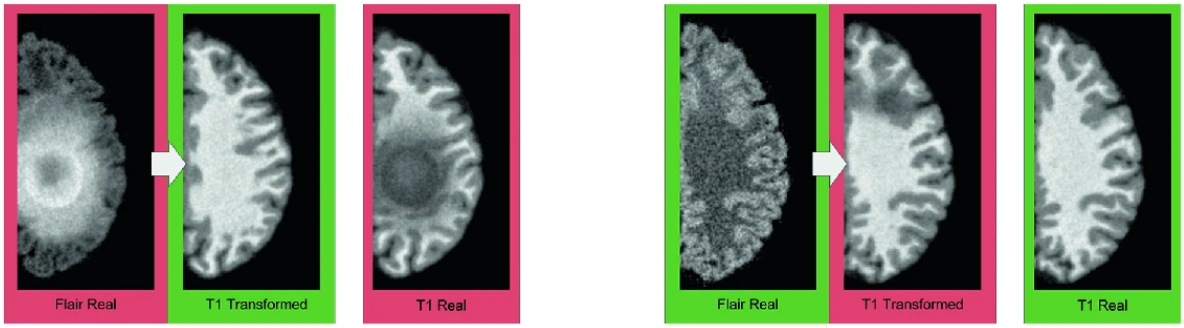


Figure 1.3: Examples of hallucinations (highlighted with green arrows) in DGM-generated images from various imaging modalities. Row 1: Incorrect number of optical disks (> 1) in images of the eye fundus. Row 2: Unrealistic texture imposed by the DGM in histology images. Row 3: Badly formed and nonsensically placed support devices in DGM-generated chest radiographs. Reprinted after cropping from Müller-Franzes et al. (2023) [34] under the Creative Commons Attribution 4.0 International License http://creativecommons.org/licenses/by/4.0/.

or removed in the generated T1 images, even if that was not the case in the original FLAIR images (refer Figure 1.4). This insertion or removal is a direct result of the tumor-present fraction in the training dataset of T1 images. Thus, the authors warn that naive image translation methods can lead to misdiagnosis, and the use of these methods might be highly dangerous for patients.

Besides demonstrations of hallucinations in biomedical images, several works on natural images have also reported artifacts that would be relevant to biomedical imaging. One kind



(a) A translation removing tumors (b) A translation adding tumors

Figure 1.4: Examples of hallucinations: unexpected tumor insertion or removal in DGM-generated images. In a domain transfer task (MRI FLAIR to T1 contrast), a DGM removed an existing tumor (left), or inserted a tumor in the DGM-generated T1 images after transforming the original FLAIR images. Both effects were a result of the composition of the training data, and the loss function employed in DGM training. Reproduced from Cohen et al. Copyright (2018) [35], with permission from Springer Nature.

of artifacts, reported by several works relates to the frequency content in the generated images. Image features occur at various length-scales, small features are represented in the high-frequency region while large features are represented in the low-frequency region in the frequency domain. Several works have reported that DGMs (i) do not always maintain the expected information at all frequencies [36–38], (ii) may not accurately generate high-frequency information [38] and information at low magnitude frequencies [37], and (iii) may be steered to maintain fidelity at prescribed frequencies [37]. Within biomedical imaging, this frequency bias can translate to only features of some length-scales being correct, while features at other length-scales being more likely to be wrong. For example, if high-frequency information is not well-learnt, small structures such as tumors and lesions might demonstrate more hallucinations as compared to large anatomical structures. Frequency bias in generation can also impact the reproducibility of texture, and any texture-based decision support system, including DNNs. Thus, a frequency bias in the generated images could also lead to mis-trained classifiers and pose risks to patient well-being.

As seen from these demonstrations, various kinds of hallucinations can occur in DGM-generated biomedical images. These hallucinations can be severe enough to cause misdiagnosis [35], yet, very few works have explored the diagnostic impacts of hallucinations in biomedical imaging [21, 39, 40]. The tremendous pace of development of DGM-based methods in biomedical imaging [2, 3, 41], together with their capacity for hallucinations that could

jeopardize patient health, have led to an urgent need for the evaluation of DGMs in biomedical imaging.

1.2 Challenges to DGM Evaluation in Biomedical Imaging

Evaluation of DGMs relevant to biomedical imaging faces some unique challenges as compared to natural images or other domains.

- Biomedical imaging is a high-stakes domain. Error tolerance in biomedical imaging is minimal, and mistakes in DGM-generated image ensembles could be harmful.
- Even visual evaluation of biomedical images requires expert knowledge [42, 43]. When novel DGMs are prototyped or existing DGMs are adapted for use with medical images, often non-clinicians train at least tens of networks on medical images based on visual feedback or tracking metrics designed for natural images. Non-clinicians will not recognize errors in anatomy and physiology, and these errors could propagate and amplify the risks in downstream decision-making. Thus, there is a need for datasets and evaluation methods that can aid the design/ adaptation of DGMs for biomedical images by non-clinicians at an early stage of method development.
- Assessing the trustworthiness of DGM-generated images in medical imaging is aggravated by the lack of a mathematically defined ground truth—e.g., there are no known statistics that reliably indicate when a generated heart is shaped realistically.
- Although several methods of evaluation are proposed for natural scenes [44, 45], and are commonly employed for evaluating medical image ensembles, their applicability is not established. (Refer chapter 2 for a detailed discussion of current evaluation methods.) In addition, an appropriate "universal" feature space for medical images is not established, unlike for natural images (e.g., the ImageNet feature space [46]).
- Ensemble measures are insufficient, and per-image evaluation of the generated images is necessary. Often, for a medical dataset, each image contains a similar set of anatomical structures with known prevalences and relative positions. These attributes have to be correct in *each image*, and not just on average over the entire image ensemble.

Thus, to realize the potential benefits of DGMs in biomedical imaging, it is essential to develop robust, domain-, and task-specific evaluation frameworks at multiple stages of technical development and clinical deployment. Most importantly, the development of benchmarks has to at least match the pace of, and ideally, be prioritized over method development to translate DGMs to labs and clinics.

1.3 Synthetic Data as a Feasible Approach to DGM Evaluation

Given the constraints described above, one feasible solution for evaluating DGMs in biomedical imaging is to design a known "ground truth", i.e., design synthetic datasets. Synthetic data could at least partially alleviate some of the challenges described in the previous section and has been employed in some works [47, 48] (detailed discussion in chapter 2). However, the potential of synthetic-data-based methods has not been fully explored for the evaluation of DGMs in biomedical imaging. The present thesis focuses on the development of synthetic-data-based methods for DGM evaluation in biomedical imaging; some advantages of this approach are discussed below.

A major advantage of employing synthetic data for DGM evaluation is the creation and availability of a ground truth. Once the "right answer" is known via the ground truth, comparisons can be made with the DGM-generated images to determine the fidelity of the generated data. Because we have complete control over the design of the synthetic data, various domain-relevant features can be encoded and assessed. Domain-relevant features could include general attributes such as per-image feature prevalence, or specific anatomical attributes such as tumors of prescribed shape and size. That is, the complexity and the realism of the data can be determined by the user, and several domain-relevant benchmarks can be established. In addition to the complexity of images, the composition of the training dataset could also be systematically varied to assess its impact on decision support. For example, the prevalence of a class (equivalent to a pathology) in the synthetic dataset could be varied to assess the minimum number of images required to generate realistic images from that class.

Synthetic-data-based evaluation enables the design of objective and automated evaluation of DGMs. Objective evaluation of biomedical images eliminates the subjectivity in human judgements, and provides a more reliable way of mitigating risks in DGM deployment. Automated evaluation proves beneficial when several tens of thousands of images have to be assessed from at least tens of DGM trainings involved in DGM adaptation; clinicians could not possibly look through tens of thousands of images only to assess one DGM.

Synthetic data can provide model-agnostic benchmarks. Therefore, the comparison of substantially different DGMs is feasible. In addition, new and emerging DGMs can be compared against the current state-of-the-art DGMs.

Of course, the evaluation of technology requires benchmarks at several stages from design to deployment. Benchmarks are different for the different stakeholders at each stage. Synthetic-data-based approaches provide the flexibility to design assessments at all stages. Early stage evaluation datasets could be designed to have high interpretability by non-clinicians whereas datasets closer to deployment could have higher realism and be designed with inputs from clinical experts.

1.4 Some Desired Properties of Generated Image Ensembles

Besides the design of the training data, another important aspect of a DGM evaluation framework based on synthetic data is to determine what constitutes a "good" generated ensemble. Some desirable qualities of a generated image ensemble are presented below.

- **Diagnostic value:** Ideally, the generated image ensemble and the training ensemble would have equivalent diagnostic value. That is, if the generated image ensemble were employed instead of the original training dataset, the downstream visual diagnosis, or clinical decision-making should not be negatively impacted. "Diagnostic value" is difficult (or potentially impossible) to capture in a single number, and hence, many different aspects of the generated images need to be tested in order to ascertain that diagnostic value is not lost.

- **Presence of certain features:** Features characteristic to the object or condition being studied and the imaging modality should be present. These features may be quantitative (including texture), positional, or morphological. Furthermore, these features must be simultaneously accurate at *multiple length-scales*. E.g., small structures such as lesions and tumors, large anatomical structures, as well as typical intensity ranges of these structures, should all be correct for a generated image to be correct.

- **Absence of certain features:** Unexpected features should be absent. That is, well-formed features but uncharacteristic of a condition, as well as artifacts imposed by a model, both should be absent.

- **Domain fidelity:** All features should respect the rules of a domain; generally, these may be known to domain experts. These rules may manifest as conditional co-occurrences amongst features and determine the "composition" of an image. Domain fidelity also includes respecting the physical attributes of an imaging system. E.g., a generated chest radiograph may contain a well-formed pacemaker, but if the pacemaker appears far away from the heart, domain fidelity is violated.

- **Ensemble level accuracy of features:** Ensemble distributions of features from the generated ensemble should match those from the training ensemble. Several popular measures of assessing DGMs for natural image generation test only this aspect, e.g., Fréchet Inception Distance [45], Inception Score [49]. E.g., over a large ensemble of chest radiographs, the approximate size of the heart or the texture of the heart should be the same.

- **Instance-level, or per-image accuracy:** Biomedical images have per-image constraints on features that may not be captured in ensemble-level measures. Testing per-image accuracy enables the identification of anomalous generation, which should be excluded from the ensemble before it is employed for any downstream use. E.g., in a large ensemble of chest radiographs, *each* generated image should have exactly the same number of rib-pairs, similar to the training data.

- **Absence of memorization:** DGMs are not expected to overfit the training data, and their intended use often demands stochastic variation in the generated ensemble. Measures testing image quality or fidelity may not capture memorization and hence, explicit testing of memorization may be necessary. E.g., if every single DGM-generated image in a large generated ensemble was an exact copy of one training example, most

measures of image quality would report that the generated image ensemble is highly similar in quality to the training data.

- **Diversity, within class, and per-class:** Assessment of the sufficiency of stochastic variation within a class, and across classes or modes in the data is required, especially because some DGMs are known to preferentially generate only a subset of the modes in the data when employed for unconditional synthesis [50]. E.g., if the training dataset represents images four different pathologies, but only two of those pathologies are always generated in a DGM-generated ensemble, the generated ensemble is much less diverse than the original ensemble even though each generated image may be perfect. Assessing memorization and diversity, both require an appropriate feature space, and hence, existing measures designed for natural images may not directly translate to biomedical images.

Assessing all these attributes may not be possible via a single number or a figure-of-merit. However, various tests can be designed that employ synthetic data to assess some of these aspects. This is extensively discussed in the next chapter.

1.5 Overview of the Thesis

A technical introduction to DGMs and an overview of the existing methods of evaluating DGMs are provided in chapter 2. Background for the concepts of "similarity" and "spatial context", which form the basis of the evaluation frameworks proposed in this thesis, is provided in chapter 2 as well.

The first of two evaluation frameworks proposed in this thesis—designed to test the capacity of DGMs to hallucinate *prescribed* spatial context—is discussed in chapter 3. This network-agnostic, data-driven framework is based on designing stochastic models of context, without explicitly modeling anatomy. It enables ruling out DGMs that do not have the capacity for certain downstream tasks.

The second evaluation framework is based on a stochastic model of anatomy and is designed to test the capacity of DGMs to hallucinate spatial context that emerges from the stochastic interactions of distinct anatomical structures. This is discussed in chapter 4. This

chapter also includes a study of several kinds of artifacts identified from a variety of DGMs benchmarked on the same dataset.

In chapter 5, the two frameworks described earlier are employed to test diffusion generative models, a kind of DGM that has recently gained great popularity due to high visual quality of generated images. A specific diffusion generative model: denoising diffusion probabilistic model (DDPM) [51, 52] is assessed for hallucinations in spatial context. The results are reported in this chapter and inferences are drawn about the potential use cases of these DGMs.

In the final chapter (chapter 6) of this thesis, the major findings are summarized and their implications are discussed within the context of biomedical imaging. Some broad directions for further exploration are also highlighted given the current state of the field.

Chapter 2

Background for the Evaluation of DGMs in Biomedical Imaging

"How it is we have so much information, but know so little?" - Noam Chomsky

2.1 Overview

This chapter provides the technical background for generative models of images, followed by an overview of the existing evaluation strategies for assessing DGMs. Both aspects: the design of new DGMs, and the design of evaluation measures, are closely related. Knowledge of DGM approaches exposes potential limitations and informs the design of evaluation approaches, whereas well-designed evaluation measures aid the protoyping of novel DGMs and determine their applicability. Following the review of DGMs and DGM evaluation measures, the concepts of similarity and context are discussed; these concepts form the basis for designing the evaluation frameworks proposed in this thesis.

2.2 A Technical Overview of Generative Models of Images

2.2.1 Early Generative Models of Images

Although the term "generative models" has become popular only recently, various methods of image synthesis have been developed over the last few decades, much before the development

of deep generative models. Early approaches of image synthesis involved generating random binary patterns with specified statistics [53], with goals ranging from understanding the relation between perception and statistics [53] (see Figure 2.1), to characterizing random phenomenon [54]. Later approaches to image synthesis aimed to generate more complex textures in grayscale or color. These texture synthesis approaches spanned a variety of subtasks such as inpainting [55], image extrapolation [56], and whole-image synthesis [56, 57, 57–59] from a single image, or an ensemble comprising several images.

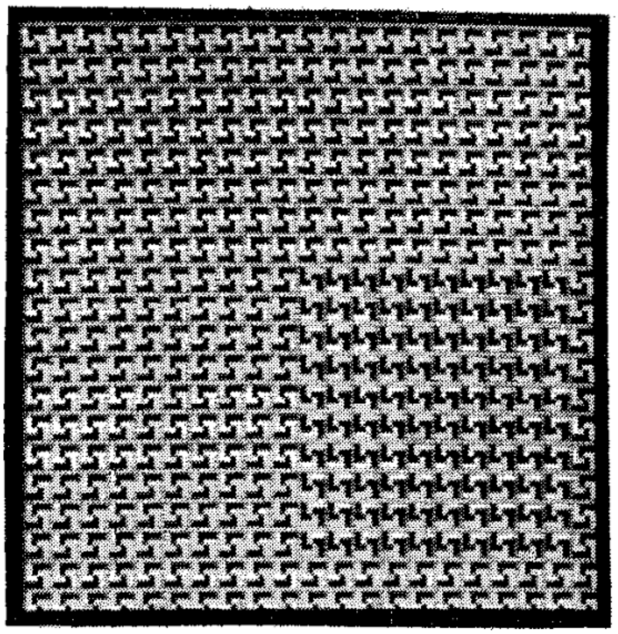


Figure 2.1: An image generated from a stochastic generative model, not based on deep learning, to study the relation of visual perception with image statistics by Julesz [53] ©1962 IEEE. This image has two kinds of textures. Note the difference in texture in the lower right quadrant of the image, obtained by taking the complement of the prescribed texture in the remaining image.

A large body of literature exists in this field and describes various parametric or non-parametric methods of image representation; some of these methods include probabilistic representation of images via Markov Random Fields [56, 57], image transforms [58, 59], and, much later, convolutional neural networks (CNNs) [60].

Of particular note was a non-parametric method that established a correspondence between information theoretical concepts originally developed for language to the composition of natural images [56]. This method defined pixel values as conditional distributions obtained

from their respective neighborhoods. In the present thesis, it is in this sense that biomedical images are understood and the concept of "context" in images is referred to. For example, if a part of a chest radiograph is missing, the missing region can be inpainted as a region of the heart if the neighborhood consists of pixels belonging to the heart, or as a background regions if the neighborhood consists of background pixels. A similar understanding of images from a visual cognition perspective was provided by Oliva and Torralba [61] for object recognition.

Although conventional methods are now being replaced by deep generative models for image synthesis applications when feasible, knowledge about image composition, textures, i.e., local intensity-derived statistics, and information content in images that was discovered via early methods of image synthesis remains invaluable for designing tools to evaluate DGM-generated images.

2.2.2 Deep Generative Models of Images

Deep generative modeling of images has enabled the generation of highly complex images with excellent visual quality. Different approaches of image representation and learning have been employed by modern DGMs for image generation while balancing trade-offs in image quality, diversity, and compute requirements [50,62]. In general, the goal of generative modeling is to learn to produce samples $\tilde{\mathbf{x}} \sim p_\theta(\tilde{\mathbf{x}})$, from the distribution of the training dataset $\mathbf{x} \sim p(\mathbf{x})$, where θ represents the parameters learned by a DGM during training. Here, $\mathbf{x}$ represents an image and the dimensionality of $\mathbf{x}$ is equivalent to the size of the flattened image. E.g., an image of size 8×8 can be represented as 64-dimensional vector $\mathbf{x} \in \mathbb{R}^{64}$. The modeling of the density of the training data may be implicit, i.e, learning to draw the "correct" random variates from a high-dimensional distribution, or explicit, i.e., accurately modeling the high-dimensional distribution itself. Amongst modern DGMs families, the target data distribution is modeled (i) implicitly by generative adversarial networks (GAN) [63], (ii) explicitly in some normalizing flow-based models [64], and (iii) approximated by variational auto-encoders (VAEs) [65] and diffusion generative models [66]. Many of these approaches are latent-based approaches, i.e., the training data is represented in a compressed or lower dimensional space than that of the training data, termed as "latent space". A trained latent-based DGM can generate images given random samples from this latent space. A brief description of the major DGM approaches follows.

Generative Adversarial Networks

Generative adversarial networks (GANs) [63] have been highly popular due to the high visual quality of images and fast sampling. GANs can be understood as two adversarial networks trained together as a min-max game. One network: the generator (G) aims to create an image from a lower dimensional latent vector $\mathbf{z} \sim p_z(\mathbf{z})$, i.e., $G : \mathbb{R}^m \to \mathbb{R}^n$, where $m < n$, and m, n, respectively represent the dimensionality of the latent space, and that of the original image. The second network: the discriminator (D) aims to distinguish images produced by G from the original training images, i.e., $D : \mathbb{R}^n \to [0, 1]$, by estimating the probability that an image belongs to the training data distribution. Both networks initially demonstrate poor performance. Over several iterations, the performance of both networks improves and the trained generator G can generate samples indistinguishable from the training dataset by the discriminator D.

The loss function for a generic GAN can be written as:

$$\min_G \max_D \mathcal{L} = \mathbb{E}_{\mathbf{x}\sim p(\mathbf{x})} \ln D(\mathbf{x}) + \mathbb{E}_{\mathbf{z}\sim p_z(\mathbf{z})} \ln (1 - D(G(\mathbf{z}))), \tag{2.1}$$

where $\mathbb{E}_{\mathbf{x}\sim p(\mathbf{x})}$ is the expectation over the training data, and $\mathbb{E}_{\mathbf{z}\sim p_z(\mathbf{z})}$ is the expectation over the distribution of the latent vectors. Several issues with training stability such as vanishing gradients and mode collapse have been observed with GANs [62,67]. Here, "mode" is understood in a statistical sense, i.e., as a measure of central tendency in a dataset. This has motivated modified versions of the loss function above (Equation 2.1). Some examples are Wasserstein GAN (WGAN), and least squares GAN (LSGAN) [68,69]. Other modifications, e.g., spectral normalization [70], gradient penalty [71], dataset-specific projections [72], self-attention mechanisms [73] have also improved training stability and image quality.

A major advancement in GAN design came from conditioning the generated input on class identity, which avoided the collapse of data modes that corresponded to specific classes in the high-dimensional data distribution [74]. In this formulation, the generator learns to generate data given a class label (typically appended to the latent vector), and the discriminator learns to distinguish between the real and generated images from the same class.

Conditioning with image inputs [75] opened even more avenues for employing these networks. Typically in image-conditioned models based on GANs, the discriminator distinguishes real and generated images as before, whereas a generator learns to transform a given *image*

via a DNN. The requirement of labeled data or paired data in two domains, precludes the widespread replacement of unconditional GANs by conditional GANs for image synthesis. In other words, if conditioning is to be employed to ensure that a class is well-replicated in the generated image ensemble, it is first essential to label that class in the training data. Similarly, if images from one domain are to be transformed to another domain, matched data must be acquired or annotated in those domains. The labeling process may involve substantial human effort or a highly robust automated method.

Variational Autoencoders

Unlike GANs, which implicitly model the training data distribution, variational autoencoders (VAEs) [65] approximate the density of the training data in a continuous latent space, typically represented via parameterized Gaussian distributions. VAEs are one of the earliest deep generative models, and have the advantages of good mode coverage and fast sampling. However, the generated images are typically blurred, leading to a lower image quality than that of other modern DGMs like GANs and diffusion generative models [50].

From a probabilistic perspective, the variational inference approach can be understood as follows. A generative model aims to learn and sample the data distribution $p_\theta(\mathbf{x})$, where θ represents the learned parameters of a model. A latent-based model aims to obtain a lossless latent representation of the true data $p_\theta(\mathbf{z}|\mathbf{x})$, where $\mathbf{z}$ represents a vector in the latent space. Because $p_\theta(\mathbf{z}|\mathbf{x}) = \frac{p_\theta(\mathbf{x}|\mathbf{z})p_\theta(\mathbf{z})}{p_\theta(\mathbf{x})}$, we require knowledge about the denominator: $p_\theta(\mathbf{x})$. However, $p_\theta(\mathbf{x}) = \int_{\mathbf{z}} p_\theta(\mathbf{x}|\mathbf{z})p_\theta(\mathbf{z})d\mathbf{z}$, where $p_\theta(\mathbf{z})$ is the prior. This integral is often not tractable for high-dimensional data, and hence, an approximation $q_\phi(\mathbf{z}|\mathbf{x})$ is employed instead of the true posterior $p_\theta(\mathbf{z}|\mathbf{x})$. Ideally, the approximation $q_\phi(\mathbf{z}|\mathbf{x})$ should be as similar as possible to the true posterior: $p_\theta(\mathbf{z}|\mathbf{x})$ and hence, we seek to minimize the difference between the two distributions by employing the Kullbeck-Leibler (KL) divergence for this purpose. Then, the minimization problem is formulated as:

$$\mathcal{L} = \min \mathrm{KL}(q_\phi(\mathbf{z}|\mathbf{x})||p_\theta(\mathbf{z}|\mathbf{x})). \tag{2.2}$$

This problem, in turn, can be reformulated via the evidence lower bound (ELBO) to obtain an equivalent maximization problem:

$$\mathcal{L} = \max \mathbb{E}_{\mathbf{z}\sim q_\phi(\mathbf{z}|\mathbf{x})} \log p_\theta(\mathbf{x}|\mathbf{z}) - \mathrm{KL}(q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z})), \tag{2.3}$$

where $\mathbb{E}_{\mathbf{z}\sim q_\phi(\mathbf{z}|\mathbf{x})}$ represents the expectation over the latents. In Equation 2.3, the first term is the reconstruction loss that promotes lossless representation of the training data via the latent representation. The second term aims to match the learned variational distribution of latents to the prescribed prior over the latent distribution.

Thus, VAEs consist of two DNNs trained together via variational inference. The goal of the first DNN, i.e., the probabilistic encoder $q_\phi(\mathbf{z}|\mathbf{x})$, is to obtain a meaningful and continuous latent representation of the original data, where ϕ represents learned parameters. The second DNN (or probabilistic decoder): $p_\theta(\mathbf{x}|\mathbf{z})$, aims to generate an image from a latent sample, where θ represents learned parameters. After training is complete, the decoder can be employed to generate new images from random latent samples.

Some improvements in conventional VAEs include tighter bounds on the objective [76], improving model expressivity via more complex priors [77] and hierarchical VAEs [78], and discretization of the latent space, e.g., VQ-VAE [79] to avoid collapse.

Although VAEs are not highly popular as standalone DGMs, they have been employed in conjunction with state-of-the-art DGMs to speed up the training of these state-of-the-art DGMs via dimensionality reduction. An example of this approach is employed in studies in chapter 5, and termed as a "latent diffusion model".

Energy-based Models

Energy-based models (EBMs), although one of the early DGMs, have greatly evolved and given rise to some state-of-the-art DGMs such as score-based [80], diffusion [51, 52, 81], and stochastic-differential-equation-based [82] generative models. All three formulations are closely related and can be derived from one another under certain conditions. The common theme of energy-based models is that a density function $p(\mathbf{x})$ is represented as:

$$p(\mathbf{x}) = \frac{e^{-E(\mathbf{x})}}{\int\limits_{\tilde{\mathbf{x}}\in X} e^{-E(\tilde{\mathbf{x}})}}, \tag{2.4}$$

where $E(\mathbf{x}) : \mathbb{R}^n \to \mathbb{R}$ is an energy function. This energy function is minimum when the generated distribution and the target distribution are exactly matched.

In one energy-based approach: diffusion models, a probabilistic trajectory is learned from a multivariate standard Gaussian distribution to the distribution of the data. To learn this mapping, first, noise is incrementally added to the original data $\mathbf{x}_0$ over a certain number of time steps T, according to a predetermined schedule. The image at the end of this process $\mathbf{x}_T$ is expected to be approximately Gaussian. The forward diffusion process is a chain of discrete Markov processes inspired by thermodynamics [51, 52]:

$$q(\mathbf{x}_{1:T}|\mathbf{x}_0) = \prod_{t=1}^{T} q(\mathbf{x}_t|\mathbf{x}_{t-1}), \tag{2.5}$$

$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I}), \tag{2.6}$$

where β_t is the coefficient of the noise variance at step t.

The reverse process is a successive denoising process to approximate $p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)$, via a loss function similar to the one in Equation 2.2.2, essentially a modified version of ELBO. Over several such forward and reverse passes, the model learns to denoise data at various noise levels, and in effect, learns a Markov chain of probabilistic transitions from noise to a sample from the training distribution. A more detailed background of the diffusion model in particular is provided in chapter 5.

Another popular approach in energy-based modeling is score-based generative models. Score-based models are similar in principle to diffusion generative models, but aim to estimate the gradient of the data distribution, i.e., a "score", instead of estimating the distribution. Estimation of the score (i) can be learnt via a successive denoising process (and other methods), (ii) is sufficient to generate samples from the distribution via a kind of Markov Chain Monte Carlo (MCMC) sampling method employing Langevin dynamics, and (iii) does not require knowledge of the normalizing denominator in an energy function (Equation 2.4).

The "score" function is defined as: $s(\mathbf{x}) = \nabla_{\mathbf{x}} \ln p(\mathbf{x})$. A score-based model seeks to match the score of the model: $s_\theta(\mathbf{x})$, with the score of the data: $s_d(\mathbf{x})$, by minimizing the Fisher divergence between the two:

$$\mathcal{L} = \frac{1}{2}\mathbb{E}_{p_d(\mathbf{x})}||s_\theta(\mathbf{x}) - s_d(\mathbf{x})||_2^2, \tag{2.7}$$

where $\mathbb{E}_{p_d(\mathbf{x})}$ is the expectation over the training distribution.

As the distribution of the data is rarely uniform, the derivative of the distribution may not be estimable in low-density regions. One way to circumvent this issue is to add noise to the data and denoise the corrupted data to approximate the score of the training data [80].

Then, the loss function in Equation 2.7 can be rewritten as:

$$\mathcal{L} = \frac{1}{2}\mathbb{E}_{p_d(\mathbf{x})}\mathbb{E}_{\tilde{\mathbf{x}}\sim\mathcal{N}(\mathbf{x},\sigma^2\mathbf{I})}\|s_\theta(\tilde{\mathbf{x}}) + \frac{\tilde{\mathbf{x}} - \mathbf{x}}{\sigma^2}\|_2^2, \tag{2.8}$$

where $\mathbb{E}_{p_d(\mathbf{x})}$ is the expectation over the training distribution, $\mathbb{E}_{\tilde{\mathbf{x}}\sim\mathcal{N}(\mathbf{x}\sigma^2\mathbf{I})}$ is the expectation over the corrupted data, $\mathcal{N}$ indicates a multivariate Gaussian distribution, σ represents the standard deviation of Gaussian noise at a certain noise level, and $\mathbf{I}$ is the identity matrix. In the argument of the expectations, the first term represents the score estimated by a model, and the second term represents the score of the data corrupted at the noise level σ. This process is undertaken at multiple levels of Gaussian noise and a network learns to denoise the data at each noise level successively, i.e., until the noise level is minimum and we converge to an estimate of the true data distribution.

Although the modern energy-based DGMs described above yield high image quality and excellent mode coverage, sampling is computationally expensive. Hence, some approaches aim to blend score-based or diffusion models with other DGM approaches such as VAEs [83] or GANs [50, 84], often at the cost of image quality.

Normalizing Flows

Normalizing flows can enable the exact computation of the data likelihood. Furthermore, their invertibility can aid in obtaining a semantically meaningful latent space representation. Normalizing flows can be understood as a chain of invertible functions. When a random variable $\mathbf{x} \sim p(\mathbf{x})$ is transformed via the action of a smooth, invertible function $f : \mathbb{R}^d \to \mathbb{R}^d$,

the output $\mathbf{y} = f(\mathbf{x})$ can be obtained via successive application of the change of variables rule:

$$p(\mathbf{y}) = p(\mathbf{x}) \left| \det \frac{\partial f^{-1}}{\partial \mathbf{y}} \right| = p(\mathbf{x}) \left| \det \frac{\partial f}{\partial \mathbf{x}} \right|^{-1}. \tag{2.9}$$

That is, a chain of K such invertible transformations f_k acting on a random variable $\mathbf{x}_0 \sim p_0$ yields:

$$\mathbf{x}_K = f_K \circ \cdots \circ f_2 \circ f_1(\mathbf{x}_0), \tag{2.10}$$

and the corresponding density $p_K(\mathbf{x}_K)$:

$$\ln p_K(\mathbf{x}_K) = \ln p_0(\mathbf{x}_0) - \Sigma_{k=1}^{K} \ln \left| \det \frac{\partial f_k}{\partial \mathbf{x}_{k-1}} \right|. \tag{2.11}$$

Major innovations have aimed to improve (i) the expressivity of the invertible functions, e.g., via coupling [85, 86], autoregressive layers [87], low-rank representations [88], and (ii) stochastic estimation of the Jacobian determinant e.g., FFJORD [89], residual flows [90, 91].

Normalizing flows provide fast sampling and good mode coverage [50], but their generated image quality is generally inferior to GANs and modern energy-based generative models. Hence, normalizing flow-based methods are relatively less common than GANs or modern energy-based approaches in medical imaging applications [92].

From the overview of DGM approaches, we see that substantial innovation and progress has occurred in the design of DGMs. However, DGMs still yield imperfect images, either as visually low-quality images, or as hallucinations in visually high-quality images. This further underlines the need for evaluating DGMs from the perspective of improving DGM design as well as assessing task-specific applicability.

2.3 Current Evaluation Methods for Deep Generative Models of Images

Many different kinds of measures have been proposed for evaluating DGMs, but only a few certain measures are typically reported in most DGM-based methods. Furthermore, each measure focuses on some aspect of a DGM and no single measure is a stand-alone indicator of DGM performance. A brief description of the major evaluation approaches are provided below; other measures exist but are not as common [44,93]. Almost all measures were originally developed for evaluating natural images, a few of these measures have been employed for biomedical images. Note that in literature, these measures are often referred to as "metrics" even though they may not be metrics in a mathematical sense (see subsection 2.4.1).

2.3.1 Ensemble-based Measures

Ensemble-based measures are based on the computation of a distance metric between two distributions obtained by projecting two image ensembles in a certain feature space [94]. That is, each image in an ensemble forms one data point in a chosen feature space and an image ensemble corresponds to a point cloud. The distance between the two point clouds corresponding to two image ensembles (e.g., training ensemble and generated ensemble) is reported as a measure of image quality. By far, this is the most popular approach of reporting the image quality of generated images, even though it does not assess the quality of individual images. Within ensemble-based measures, the most popular method is the Fréchet Inception Distance (FID), [45] which computes the Fréchet distance between two distributions corresponding to the training and generated datasets, after each of the two datasets are projected into a feature space (typically from a pre-trained natural image classifier such as Inception v3 [95]), and fit with a multidimensional Gaussian distribution. Several versions of the FID score have been proposed to: standardize the method [96], make its computation more efficient [97], employ more distinct spatial features [98], or remove/ reduce bias in the estimation [99]. Essentially, the FID captures the first two moments of the feature distribution extracted by a pretrained classifier, and its relation to human perception may vary depending on the use case [100]. Although the FID is also commonly employed to evaluate generated medical images, the relevance of the natural image feature space to medical images has not yet been established, to the best of our knowledge. There are at least two

major problems with FID as a measure of evaluation. First, multiple ensembles could have the same FID score but demonstrate different kinds of errors, obvious to human observers, e.g., one image ensemble could have small smudges in all images, another ensemble could have minor discontinuities at edges and no smudges, however, both ensembles could have the exact same FID score. Second, the change of FID score cannot be related to specific errors in images, e.g., if one image ensemble has an FID score of 2, and another ensemble has an FID score of 4, no obvious inference can be made about the kinds of errors in the second ensemble as compared to the first.

Another popular measure is the Inception Score (IS) [49], which reflects the diversity as well as the quality of a generated dataset. This score is computed only on the generated data and excludes any comparison to real images. Specifically, to obtain this score, an Inception model, i.e., a classifier pretrained on the ImageNet dataset [46], is employed to predict class label probabilities on each generated image. A "good" IS is one that minimizes the entropy of conditional probabilities and maximizes that of marginal probabilities. The IS also suffers from the issue of unclear relevance of its feature space for medical images, similar to the FID.

Kernel-based ensemble methods such as the maximum mean discrepancy (MMD) [101], which is particularly popular for graph neural networks [102], and kernel inception distance (KID) [103] have also been proposed for natural images, but are not as commonly reported for medical images as the FID and IS.

Within medical imaging, another ensemble-based approach involves observer studies [104]; these studies may employ human observers [42,105,106], e.g., in two-alternative forced choice tests, or numerical observers [106, 107], e.g., Hotelling observer, and the Bayesian ideal observer, for assessing the acceptability of generated image ensembles with respect to the training image ensemble. Human-observer-based studies typically require domain experts to visually assess the generated images, which may not always be practical at the early stages of technology development. Numerical observers, based on Bayesian statistical decision theory, do not suffer from this limitation and may provide valuable information about the efficacy of DGMs for a specific task. However, they may not provide human-interpretable information about the kinds of errors made by DGMs within each image.

2.3.2 Spectral Methods

Recall that a frequency-domain representation of images, i.e., a Fourier transform of the images, enables assessing image features according to their size or length-scale. Small features such as tumors are represented in the high-frequency region, whereas large features such as the heart are represented in the low-frequency region. Errors in specific frequency ranges can thus be related to errors in image features at certain length-scales or sizes.

Methods in the frequency domain have been developed not only to investigate model bias [37, 108], but also have been extensively employed for deepfake detection [109–112]. Frequency bias in generation as well as network artifacts manifesting as unexpected information at certain frequencies can both be captured via spectral evaluation. Multiple works [37, 38] have shown that learning high frequency information is not an easy task for the GAN family of DGMs, and that high-frequency artifacts are common in GAN-generated images. One of these works [38] has also demonstrated that learning can be steered to maintain fidelity in a specified range of frequencies. Thus, it is possible that frequency bias and high-frequency artifacts may impact the diagnostic value of biomedical images. However, a task-specific analysis of the impact of frequency bias on medical imaging tasks remains to be performed. Furthermore, these findings suggest that a complete evaluation of DGM capacity for biomedical image synthesis must include measures that account for multi-scale correctness of spatial features, going beyond low-order spatial statistics.

2.3.3 General Measures of Image Quality

Some established methods of comparing the image quality of two or more images are also employed for DGM evaluation in some cases, as applicable. These include peak signal-to-noise ratio, contrast-to-noise ratio, and structural similarity index measure [113]. Although these measures are commonly employed in practice, even with medical images, they represent technical efficacy and may not represent the diagnostic value of the images [114–116]. Furthermore, computation of some of these measures require knowledge of the ground truth, which likely is not available. General measures of image quality are typically employed for image-conditioned DGMs, when data in two domains is available, e.g., segmentation task, domain transfer task. These measures may not be directly applicable for unconditional image synthesis.

2.3.4 Measures Based on the Data Manifold

To improve the interpretability and potentially the steerability of DGMs, data manifold methods have also emerged as tools for evaluating DGMs. Two popular measures aim to identify a data manifold that that is also perceptually relevant: perceptual path length (PPL) [117], and learned perceptual image path similarity (LPIPS) [118]. PPL has been shown to be superior to FID in terms of understanding image composition and perceptually relevant differences in image quality [119]. LPIPS has been employed in data poisoning tools to protect copyright [120] by inducing imperceptible perturbations in the original images, making them useless as potential training data for generative models.

Another approach involves the disentanglement of image attributes in the latent space of a DGM [121–125]. An example is the intrinsic multi-scale distance (IMD) [122], which is a multi-scale measure that takes into account the intrinsic structure of the data as well as all the moments of its distribution. Similar measures that explore the topology of the latent manifold only do so at a global scale [123, 124]. In the medical imaging domain, disentanglement of the latent space has served several purposes [125], ranging from disentangling anatomical factors to predicting treatment response. Evaluation measures based on data manifolds have the advantage of being DGM-agnostic and to some extent, explainable, but may also vary in their results based on changes in the composition of a dataset. One limitation of data manifold methods is that the estimation of the manifold itself might be dependent on the choice of hyperparameters, the number of samples in a dataset, and the diversity in the training data itself. That is, evaluation results might change if a dataset was halved in size, or some classes were not well-represented in the data.

2.3.5 Evaluations via Synthetic Data

Synthetic data is a powerful tool to explore the capacity of DGMs; however, as compared to other evaluation methods, this approach has received less attention. In this approach, typically, a dataset is designed to test specific capacities of a DGM, after training the DGM on this dataset. The performance of the DGM is determined based on the composition of the DGM-generated ensemble. One work [126] employed several explicitly parameterized distributions to train DGMs and studied the impacts of dimensionality, size and complexity of the training dataset, and robustness to hyperparameters. Similarly, another work [127]

employed categorical data, binning a high-dimensional probability space with varying levels of coarseness, and then casting the problem of DGM evaluation as a statistical identity testing problem. Other notable works have created synthetic datasets to explore overfitting in multiple modern GANs [128], and mode collapse [129]. Synthetic data has also been employed to assess the compositional ability of diffusion generative models [130]. In the domain of large language models, recently a work [131] has adapted a synthetic dataset designed to test visual conceptual abstraction abilities to assess large language models. However, creation of realistic and high-dimensional synthetic datasets is highly challenging, and most works on natural images have employed simple, and/or low dimensional synthetic datasets.

In the medical imaging literature, although synthetic data or virtual phantoms have long been used to characterize imaging systems and assess technology performance [132–136], their application for the evaluation of DGMs has been relatively rare. One work [137] employed synthetic data to demonstrate that medical image statistics may not be correctly reproduced even though low FID scores were achieved in the training ensemble. The present thesis aims to realize some of the unexplored potential of synthetic data for the evaluation of DGMs in a data-driven manner.

2.3.6 Fidelity and Diversity Measures

Precision and recall [138] were first proposed as evaluation measures for DGMs to assess the fidelity and the diversity of the generated data. In these measures, two regions corresponding to the training dataset and the generated dataset are established in a feature space, and the overlap in these regions are compared. Precision measures how similar the generated data is with respect to the training data, while recall measures how much of the training distribution was captured in the generated data. Thus, for a DGM-generated dataset, precision can represent its fidelity/ quality and recall can capture missing modes in the data distribution. Some drawbacks of precision and recall measures were identified such as lack of robustness to outliers [47, 139], and the inability to detect identical distributions [139, 140]. To alleviate these issues, some subsequent works introduced measures based on precision and recall such as precision-recall curves instead of a single number [140], definition of precision-recall that was robust to outliers, along with a new measure for generalization [47], and novel measures e.g., density and coverage, which are based on refined definitions of data neighborhood and overlap computation [139]. Adapting precision and recall measures for medical imaging data

requires a relevant feature space to be defined. Some works in medical imaging literature [34,40] report precision and recall, (in addition to FID) to demonstrate quality and diversity of the generated image ensembles.

Although fidelity and diversity measures provide important information about the composition and diversity of a DGM-generated image ensemble, fidelity measures still suffer from the same issue as ensemble-based measures because each image is still a data point in a potentially high-dimensional feature-space.

2.3.7 Model-specific Measures

Model-specific measures have been developed to explicitly explore the generative capacity of specific architectures. However, inverting a DGM (which may not be inherently invertible by design) is not a trivial task. Implicit generative models such as GANs have been inverted to study the limits of their generation capacity and for image editing [141,142]. Conceptually, when an image is propagated through an inverted GAN, a latent encoding is obtained; propagating this latent encoding through the regular GAN generates another image, which ideally should match the first image. Missing features in the reconstructed image can indicate features that may not be represented or learnt by a GAN. Recently, other state-of-the-art DGMs (energy-based models) have also been inverted, but mostly to enable image editing and controlled image generation [143,144] and not for the evaluation of DGM capacity.

Although model-specific methods can identify biases in generation, these methods do not find widespread use due to their architecture-specific nature.

2.3.8 Task-based Measures

Ultimately, in a biomedical imaging workflow, a DGM is expected to be deployed for *a specific task* involving a certain imaging system. Hence, task-based measures can be valuable tests in the final stages of DGM evaluation prior to deployment; these measures may be designed in consultation with clinicians.

Typically, task-based approaches measure the performance of an observer (human or model) for a downstream task on a DGM-generated image ensemble and on the training data. If

the task-performance is equivalent on both image ensembles, the DGM-generated ensemble is considered on par with the training data. However, the performance on one task likely will not translate to performance on another task. Therefore, a DGM-generated ensemble must be evaluated for each task separately before deployment.

Often, a classification or segmentation task has been employed for task-based evaluation in the computer vision literature [145, 146]. However, even in task-based analyses, the chosen evaluation measures representative of task performance must be appropriate, e.g., the Dice similarity co-efficient may not be accurate for measuring segmentation accuracy of very small structures, or for object detection tasks involving multiple objects [147, 148]. In the medical imaging domain, relatively few works have undertaken a task-based evaluation for DGMs, although task-based evaluations employing signal detection theory are well-established in this domain [104]. A few examples of task-based DGM evaluation in medical imaging include the evaluation of the diagnostic value of DGM-generated mammography images [21], evaluation of global consistency in whole-body MRI synthesis [149], detection and estimation tasks for unconditional synthesis in multiple imaging modalities [107] and also for the evaluation of GANs [150, 151].

Task-based measures can be extremely useful for determining the ultimate acceptability of a DGM-generated ensemble, however, by principle, they may not be helpful for drawing inferences about DGM generalizability to other tasks, or per-image errors. The methods developed in this thesis are complementary to task-based measures. The proposed methods are aimed to assess the applicability of DGMs in a domain-agnostic and interpretable manner, and thus, are intended for use in the early stages of DGM adaptation.

Thus, a variety of DGM evaluation approaches exist, and each approach has its pros and cons. At the crux of each approach, the goal is to assess whether a DGM can generate a dataset similar to the training data in some sense; the definitions of similarity vary with each approach.

The following section elaborates on the concept of similarity. First, the fallacy of similarity as a distance metric is discussed. Then, an alternative approach to similarity based on set-theoretical concepts is introduced for biomedical images. This approach is then extended from the perspective of biomedical decision-making, which in turn lays the foundation for understanding the role of context in biomedical images. A general context-based evaluation framework that forms the basis of this thesis is presented at the end of this chapter.

2.4 Defining Similarity

The need for defining similarity arises from evaluation scenarios such as: are two images/image ensembles similar in quality, or is a given image (ensemble) similar to another "gold standard" image (ensemble) for some purpose. The design of evaluation measures or "metrics" is based on the notion of similarity. Typically, a similarity metric is a distance function between two representations and summarizes differences in two images (or distributions) into a single number e.g., ensemble-based measures described in the previous section. However, perceptual "similarity" is rarely, if ever, a metric [152].

2.4.1 Fallacy of a Similarity Metric

Mathematically, a metric satisfies the properties of positive-definiteness, symmetry, and the triangle inequality. Consider the process of visual diagnosis of medical images by radiologists with respect to each of the three properties of a metric.

1. Positive-definiteness: In case of ambiguity in the manifestation of a disease, different radiologists or even the same radiologist at different times may vary in their diagnosis of the same patient (typically termed as inter-reader and intra-reader variability). That is, the same stimulus may elicit different outcomes, and thus the similarity distance between an image with itself may not always be the same [153].

2. Symmetry: Second, perceptual similarity is not necessarily symmetric, and can be directional. For example, the presentation of a disease is compared to its prototype or "textbook" appearance, or to a common non-medical object [152, 154, 155], e.g., pneumonia looks like ground glass on a chest radiograph, and not vice versa.

3. Triangle inequality: Perceptual similarity may not be transitive [152]. A PET image of a brain tumor may appear similar to a brain MRI of the same patient. The latter may appear similar to a brain MRI of another patient, but that does not imply that the PET image of the original patient is similar to the MRI of the second patient.

Thus, perceptual similarity may not necessarily be a metric in clinical scenarios involving visual diagnosis.

For images in general, it has already been established that there is no linear correspondence between pixel-representations of images and changes in Euclidean distances, especially under certain image transformations [156]. The same holds from a mathematical perspective of perceptual similarity. In addition, distance measures in complex feature spaces are rarely calibrated for perceptual similarity [118].

2.4.2 A Set-Theoretical Perspective on Similarity

Although perceptual similarity may not be a metric, a set theoretical approach holds promise for quantifying similarity, as demonstrated in a landmark work by Tversky [152]. Some relevant concepts are elaborated below for medical imaging scenarios.

- Feature matching: Innumerable features could be computed from an image. But, any task involving similarity of two images can be represented via three sets of features: those unique to the first image, those unique to the second image, and those common to both images.
- Monotonicity: When comparing two images, their similarity decreases with the addition of distinctive features, but increases with the addition of common features. Thus, a monotonic scale can be established based on the number of common and distinctive features. Furthermore, depending on the task, distinctiveness or similarity of features might be weighted differently, creating a task-specific scale for representing similarity. Here, distinctive features could represent disease-specific manifestations while common features could represent non-specific manifestation of disease, or normal conditions.
- Assymmetry: Not all tasks involving similarity are symmetric. For example, a symmetric task can be one where a diagnosis has to be chosen from either disease A or disease B, both of which seem equally probable. On the other hand, testing if DGM-generated images are similar to the true medical images is a non-symmetric task, where the true data is more salient and demonstrates exemplar features.
- Diagnosticity: The importance of a feature is dependent on a task, and the set of features that it occurs within. For example, a certain radiological sign might be the distinguishing feature between disease A and disease B, and hence be the most important of all present signs/symptoms. However, the same sign might carry less importance,

if the decision has to be made between disease A and disease C, wherein this sign is often present in both diseases and thus, other symptoms might have greater diagnostic importance.

- Appropriate scale of features: Features can be identified at different scales. However, they may be diagnostically relevant only at a certain scale, and this scale must be respected for a feature set to be meaningful. For example, a radiological sign might occupy tens of pixels in standard sized chest radiograph. If all features are described at a scale of hundreds of pixels, or at a scale of single pixels, the radiological sign might lose its diagnostic value in a set of features.

2.4.3 Similarity, Identity, and Decision-making

The set-theoretical approach to similarity described above can be extended to medical diagnosis tasks via symbolic logic [157, 158]. Consider a situation when a patient presents with a set of symptoms $S_p = \{s_1, s_2, ..., s_p\}$, or a symptom complex. The clinician compares these symptoms against several diseases $D = \{D_A, D_B, ...\}$ via the corresponding disease-specific sets of symptoms, indicative of disease identity, e.g., $D_A = \{a_1, a_2, ..., a_d\}$, $D_B = \{b_1, b_2, ..., b_d\}$, where the subscripts $A, B, ...$ indicate a disease name. Based on this comparison, some diseases are ruled out, i.e, $D_{absent} = \{D_i | S_p \cap D_i = \phi\}$, for $i \in \{A, B, ...\}$, where D_{absent} indicates the set of diseases that are ruled out. The remaining set of diseases may then be explored in a probabilistic manner, such that each disease-specific symptom is also associated with a probability of occurrence. Thus, the identity of a disease and its similarity with other diseases may be represented via a set-theoretical approach as opposed to distance metrics. Last, based on a variety of factors including, but not limited to, medical knowledge, the most likely diagnosis and a recommended course of action is determined. Note that prognosis also takes into account the concepts of utility and value theory, but that is not the focus of this thesis.

2.5 Introduction to the Concept of Context

In the setup described above, a disease complex is a set of symptoms that typically co-occur at certain probabilities in the manifestation of that disease. Similarly, in biomedical images,

certain conditional co-occurrences of features, or pixel arrangements, are representative of specific, biomedical conditions. This is termed as "spatial context". Domain experts typically acquire tacit awareness of context through training and experience. Thus, context is domain knowledge which may not be obviously rule-based or learned from only one given image.

The relation of spatial context with object recognition has been demonstrated in some works from the computer vision literature [159, 160]. For example, an object such as a sheep in a natural image can be identified via the presence or absence of contextual attributes such as wool, horn, leg, head. Besides object recognition, contextual information in natural images has also been employed for tasks such as inpainting [161–163] and feature learning [164]. The cognitive role of context in visual identification of objects is also well established [61,165,166]. Antonio and Torralba [61] have demonstrated that instantaneous decisions about object identity can be made by humans based on the context of an object within an image, even when the object is not clearly visible–which is exactly the sort of thing we want DGMs to do. Thus, it is possible that context impacts visual diagnosis. In the medical imaging domain, one work [154] has recently focused on identifying classification-relevant context for explainable diagnoses on mammographic images. Yet, this idea has been relatively unexplored in the evaluation of DGMs for diagnostic imaging. In this thesis, assessment of recoverable spatial context is the central idea employed for designing evaluation frameworks for DGMs.

2.6 A General Framework for the Evaluation of DGMs Based on Reproducible Context

When deploying DGMs in any mission-critical application, it is vital to have *objective* measures of image quality [104] which are beyond subjectively "looking good" to untrained human observers. Assessing the degree to which expected spatial context is reproduced in individual images generated by modern DGMs provides one way of objectively assessing image quality. Furthermore, context-based evaluation frameworks are designed such that they may also reveal the presence (or absence) some general capacities of DGMs, rendering those DGMs inapplicable for certain downstream tasks.

The general method for both evaluation methods proposed in this thesis is as follows. First, a training dataset is purposefully created and a modern DGM is trained on this dataset.

Next, a large ensemble of images is generated from the trained DGM. Context is extracted as features from both the training and generated ensembles and, finally, the extracted features are compared to test how well the expected context was reproduced. An overview is provided in Figure 2.2.

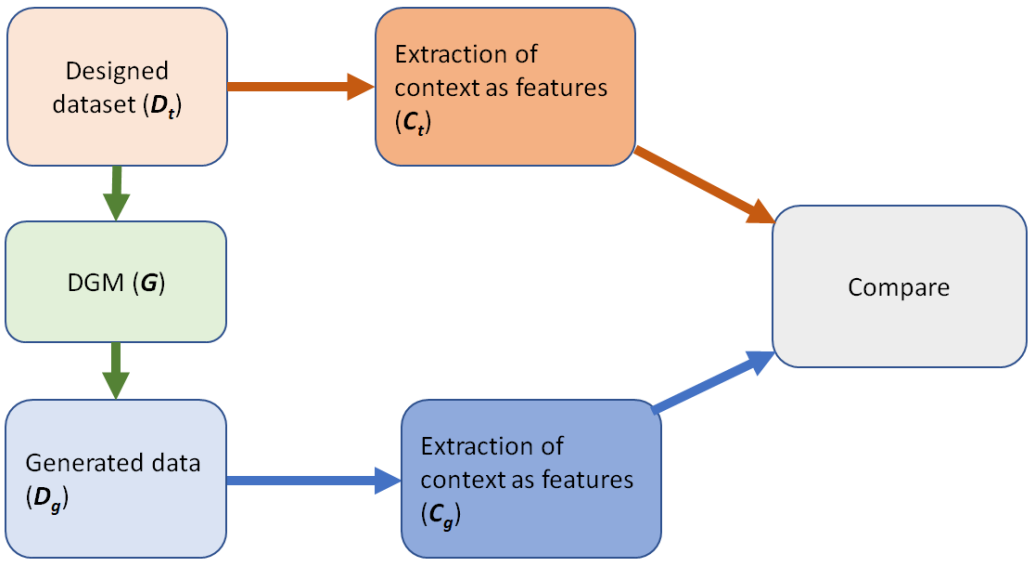


Figure 2.2: A general framework of the evaluation methods proposed in this thesis.

Two specific versions of this general framework are designed and demonstrated in the following two chapters. In the first framework, context is explicitly prescribed and encoded in the training dataset. Various contextual constraints that are relevant to biomedical images are encoded in several purposefully designed stochastic context models described in Chapter 3. It is ensured that this context should be recoverable from images generated by a successful DGM. In the final step, it is this prescribed context, along with implicitly arising context, that is tested for reproducibility.

In the second framework described in Chapter 4, a stochastic model of anatomy is adapted to create the training dataset. In this case, context is not explicitly prescribed but arises implicitly from the stochastic interactions of the different anatomical structures. Hence, in the final stage, features representing implicit context are extracted and compared.

Note that both evaluation frameworks are model agnostic.

Chapter 3

Employing Stochastic Context Models for the Evaluation of DGMs in Biomedical Imaging

"Only with respect to a projection rule, things are similar." - Ludwig Wittgenstein

3.1 Overview

The first evaluation framework based on the concept of assessing hallucinations in spatial context is presented in this chapter. The innovation lies in the creation of a synthetic training dataset that enables the purposeful encoding and recovery of domain-relevant context. Three stochastic models were designed to encode various constraints relevant to biomedical imaging and create training datasets. Although the designed models do not describe anatomy, their simplicity and interpretability enable the assessment of general DGM capacities even before they are deployed for biomedical imaging tasks. The evaluation procedure was demonstrated on two modern DGMs for the task of unconditional image synthesis and the error-rates were quantified. Several contextual errors were identified in all generated image ensembles and the implications of these errors are discussed in biomedical imaging scenarios. Last, a future direction of this work: stochastic models of context for assessing image-conditioned DGMs for image-to-image translation tasks, was briefly explored.

3.2 Introduction

Biomedical imaging applications generally require significant domain expertise. Hence, designing objective measures can be especially challenging because it usually is not clear which computable features, if any, express the knowledge of the domain expert. Therefore, a reasonable starting point toward comprehensive objective assessment of image quality, is to measure the general capacity of a DGM to reproduce sophisticated contextual features which are known prior to training.

In this chapter, the purposeful design of stochastic context models (SCMs) that encode domain-relevant, external knowledge, or "spatial context", and the use of the per-image rate of spatial context reproduction as an objective assessment of the capacity of any generative model of images is proposed. Here, spatial context may be implicit, i.e., arising from chance co-occurrence of image features, or explicit, e.g., an ineluctable pixel-placement rule defined by a human user. The design of the SCMs itself is a demonstration of spatial context being built into training images algorithmically. Each SCM can be employed to yield a large ensemble of training images for DGMs, wherein every image in this ensemble exhibits the prescribed or explicit spatial context as well as the consequentially arising implicit context. Thus, a single *generated* image will be considered useful if the prescribed context is exactly present. The role of the proposed SCMs is similar to that of stochastic object models (SOMs)—which are commonly employed in the development of imaging systems—in that each serve as a ground truth; however, there is a key difference. Here, the SCMs are generic models of a variety of task-relevant spatial contexts which can appear across a gamut of SOMs and, therefore, should not be thought of as an attempt to model any one particular object or system.

To be clearer still, it is not proposed to accurately model any particular object or image for any particular application. Instead, it is proposed to model some *kinds* of relative pixel arrangements that are generally important across many applications at once [133]. Consider a wrist radiograph of a human. There is a known number, location, and size of wrist bones *relative to* each other. Here, we are not proposing to model the wrist, but, instead, propose to model the frequency and relative location of sophisticated image features. For example, in all wrist radiography, there are typically eight unique wrist bones that have roughly fixed positions relative to each other. We do not model the bones themselves but instead test the DGM for the capacity to generate a correct number of locally sophisticated features

(e.g., wrist bones) in a correct spatial arrangement (e.g., fixed pairing). Thus, our tests go beyond any one anatomical model in that we are assaying a generic capacity to generate features under known contextual constraints. In this way, we can rule out DGMs which may not have the capacity to maintain context and, thus, in the current example, would be ill-suited for a downstream use involving the per-image bone count. Furthermore, with the same dataset, we can also rule out DGMs for other tasks in domains besides medical imaging, where per-image prevalence may be of importance. In other words, because the proposed method does not involve explicit modeling of a specific object or system, it can translate to tasks across several domains. The recoverable spatial context that is proposed to be encoded within each training datum reflects both external and high-order knowledge of correct spatial arrangements. It is external in the practical sense that what should be true about every image may not be learnable from any one image; it is high-order in the sense that correct appearance of features in any one image is not readily expressible in, or detected via, grayscale histograms or variance-covariance matrices. Therefore, it is also explicitly noted that throughout this thesis "order" should not be confused with the *degree* of moments of any particular probability distribution.

3.2.1 Overview of the Proposed Methodology

In this chapter, it is demonstrated that both implicit and explicit spatial context can be built into training images algorithmically, such that it can be verified readily after generation, and without specifying formulas for describing any particular image feature. This means that we have a ground truth for testing generated images for various contexts. Then distinct SCMs were employed in several experiments to assess the extent that DGMs learned high-order information along with whatever low-order information was also learned during training. The goal of this work is to provide a data-driven method, independent of generative model architecture, that enables the assessment of DGMs for their capacity to reproduce domain-relevant, high-order spatial context.

3.3 Methods

3.3.1 Description of the SCMs

Three families of SCMs are described in the following subsections. All realizations from all SCMs are 8-bit grayscale, 256x256-pixel images; sample realizations are shown in Figure 3.1. The three ensembles (in order of presentation) comprise 32768 (per-class), 65536 (per-class) and 131072 images respectively. Ensembles from the designed SCMs have been made available on Harvard Dataverse: https://doi.org/10.7910/DVN/HHF4AF (version 2). The post-hoc analyses codes are available on github: https://github.com/comp-imaging-sci/scms-dgm-evaluation.

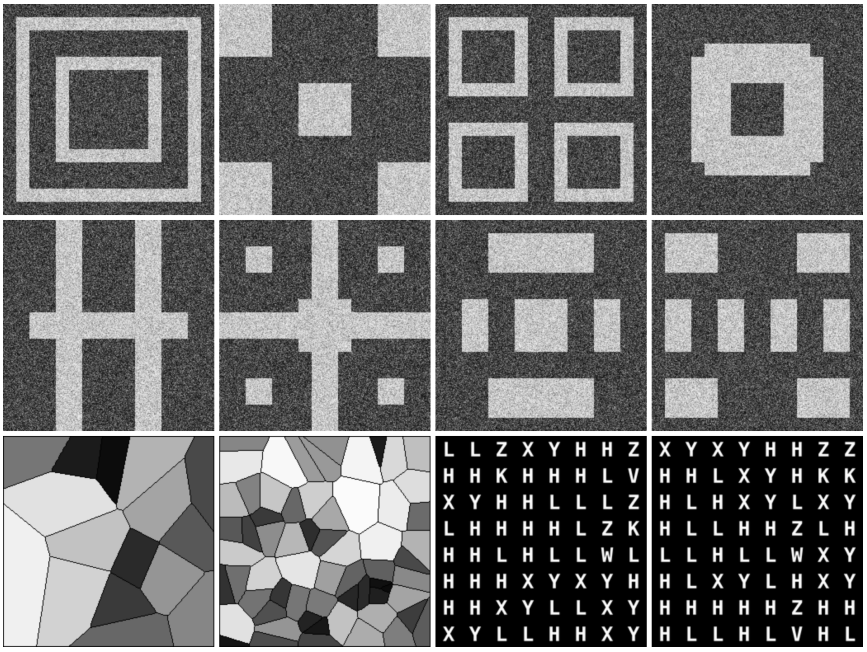


Figure 3.1: Sample realizations from the three purposefully designed SCMs. Top two rows: One realization each from the eight classes in the flags SCM. Bottom row, left to right: Realizations from the shaded Voronoi SCM representing classes 16 and 64, and the alphabet SCM are shown.

Flags SCM (F-SCM)

The eight-class flags SCM was designed for testing the joint reproducibility of pre-specified, first-, second-, and high-order image features at once. This model is intended to study class-specific features, feature-specific intensity distributions and texture, and class-distribution in the image ensemble.

Each image I in any class c, can be delineated into a regular grid of 16×16 pixel tiles with each tile corresponding to either foreground f_k or background b_k, where k is the tile index, indicating tile location within the grid. Furthermore, $I_c = \{80 \times f_k, 176 \times b_k\} \quad \forall c$; this eliminates the zero-order variance in the number of pixels of interest.

Any realization in a class can be represented as:

$$I_c = \sum_k (a_{kc} f_k + (1 - a_{kc}) b_k), \tag{3.1}$$

where $\mathcal{A} \in \{0, 1\}^{K \times C}$ is a binary matrix indicating background (0) or foreground (1) for all K tile indices in C classes. Thus, $\mathcal{A}$ indicates the prescribed, class-specific foreground patterns, and an image class is one of eight distinct foreground arrangements.

Grayscale variates within f_k and b_k were chosen from distinct Beta distributions:

$$f_k \sim 152\, X + 96, \text{ where } X \sim \text{Beta}(\alpha = 4, \beta = 2), \tag{3.2}$$

$$b_k \sim 192\, X + 8, \text{ where } X \sim \text{Beta}(\alpha = 2, \beta = 4). \tag{3.3}$$

These grayscale variates were rounded to ensure discrete values before placement. The result was an image with foreground brighter than the background, but not perfectly segmentable via a single threshold. The corresponding distributions are shown in Figure 3.2.

Moreover, the placement of the variates for f_k and b_k was completely random, that is, without any prescribed correlations in pixel locations, within each corresponding tile f_k or b_k.

Last, a set of certain 24 tile-location indices k was never part of the foreground, in any class:

$$n = \{k : a_{kc} = 0\} \quad \forall c. \tag{3.4}$$

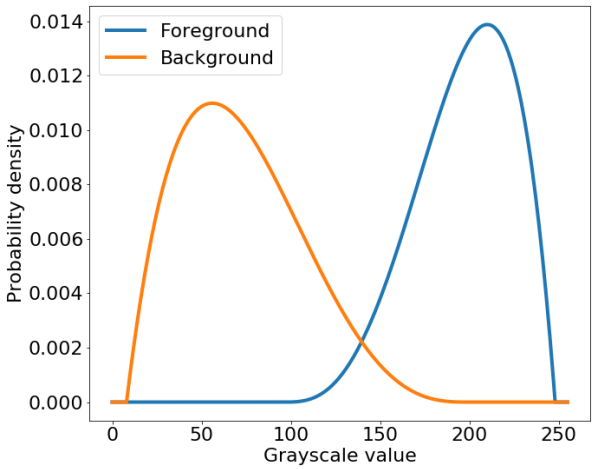


Figure 3.2: Foreground and background intensity distributions in the Flags SCM. Note that the means of the background and foreground distributions are clearly different although there is some overlap between the two.

The regions that were never foreground in any class are shown in white in Figure 3.3.

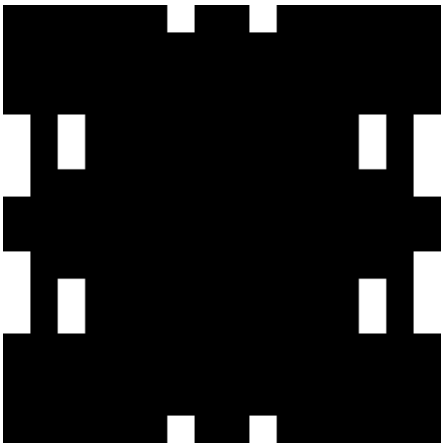


Figure 3.3: Regions forbidden as foreground in all classes of the Flags SCM are shown in white.

This is analogous to structural constraints in location for a feature. Together, these classes enable a variety of experiments for exploring how much of each informational order the DGM 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. Furthermore, the prevalence of classes in the generated ensemble can also be measured; class prevalence is one example of external, domain-specific knowledge.

The tiled nature of the images eased post-hoc classification. Each tile in I was identified as b_k or f_k by comparing its intensity mean against a threshold of 140 chosen to be halfway between the two modes of the combined grayscale distribution of b_k and f_k. The class (c) was then determined by computing the mean absolute error against each column in $\mathcal{A}$.

Voronoi SCM (V-SCM)

The Voronoi [167] SCM, a four-class SCM, enabled testing of second- and high-order information from randomized sets of image features. This SCM is intended to study per-image feature prevalences, class-specific features, and quantitative fidelity conditioned on morphology.

Each image I can be represented as a union of the set V of Voronoi regions v_i and their edges e. Here, $i = \{1, 2, ..., c\}$, where $c \in \{16, 32, 48, 64\}$ represents the cardinality of V within each I and defines the image class. Within each I, region centers were placed in a spatially random manner, unlike the fixed foreground locations in the flags SCM; this provided an additional source of object variance. Edges e were set to an intensity level of 0; this enabled robust segmentation of e and v_i from a given I. All pixels in a v_i were allocated a single grayscale value g drawn from a set of 64 predetermined, equidistant values between 8 and 255. Most importantly, the grayscale value increased monotonically with area, which is a high-order feature:

$$\rho(\text{area}(v_i), g) = 1, \tag{3.5}$$

where ρ is Spearman rank-order correlation coefficient.

In case of the *unshaded* Voronoi experiment (see subsection 3.4.2), all regions v_i were set to a grayscale value of 255. The Voronoi SCM is representative of images with multiple, positionally independent regions of interest within an image, each having a distinct intensity, e.g., histology images. The Voronoi SCM also allowed for testing the ensemble class prevalence, but with feature sets at multiple spatial scales, simultaneously. For the analysis of generated images, post-processing involved identification of the edges e, by thresholding

each I against an intensity level of 64 for the unshaded Voronoi, or via Sauvola thresholding for the shaded Voronoi, followed by "skeletonization", i.e., retaining a single pixel medial representation of all connected pixels. The skeleton was then employed for detecting v_i, which in turn determined c and enabled the extraction of region-wise values of g. It is noted that although this method of region detection is not perfect, it is still sufficiently robust for the experiments proposed. Calibration of this method on the training data predicted the mean detected number of regions exactly, with errors no greater than $\pm$ 0, 1, 1, and 2 regions for the four classes sequentially for an overwhelming majority ($> 99\%$) of the realizations.

Alphabet SCM (A-SCM)

This SCM is intended to study per-image feature prevalences, and conditional co-occurrences of per-image features.

Each realization I from this SCM can be delineated into a grid, yielding 32×32 pixel tiles t such that each t represents a letter in the alphabet $\mathbb{A} = \{H, K, L, V, W, X, Y, Z\}$. The per-realization prevalence of all letters within the image I was fixed according to the prescribed set $\mathbb{B} = \{24 \times H, 2 \times K, 16 \times L, 1 \times V, 1 \times W, 8 \times X, 8 \times Y, 4 \times Z\}$. Thus, each realization can be represented as:

$$I = \{t_{r,c} : \bigcup_{r,c} f(t_{r,c}) = \mathbb{B}\}, \tag{3.6}$$

where $f(t) : t^{32\times32} \rightarrow \mathbb{A}$ represents a template matching operation, and r, c are respectively the row and column indices of t within the grid. In other words, each image I comprises letter-tiles $t_{r,c}$ that together represent the complete set of specific letters at prescribed prevalences, i.e., $\mathbb{B}$. Although the locations of specific letters within I could vary—thus, providing random variation across realizations—they were always constrained by the following rules of conditional prevalence obeyed within each realization:

$$p(f(t_{r,c+1}) = Y | f(t_{r,c}) = X) = 1, \tag{3.7}$$

$$p(f(t_{r,c}) = Z | f(t_{r+1,c}) \in \{V, W, K\}) = 1. \tag{3.8}$$

That is, the letter Y was always preceded horizontally by the letter X (Equation 3.7), and the letters V, W, K were always preceded vertically by the letter Z (Equation 3.8). Thus,

four *ordered* letter-pairs occured in each realization: X-Y (horizontal adjacency), and Z-K, Z-V, and Z-W (vertical adjacency). Furthermore, the per-realization prevalences of the letter-pairs were fixed as 8, 2, 1 and 1 respectively. Together, these rules rules of prevalence and placements constitute the explicitly prescribed context encoded in this SCM.

For post-hoc processing, error for each t was computed as the pixel-wise difference from the known letter templates and a reasonable acceptance threshold (75% of the maximum error) was chosen once by visual inspection. Although the post-hoc classifier assigns an identity to all letters, only automatically recognizable letters were retained. This abates the effect of minor feature shape variance in further analysis.

3.3.2 Network Trainings

The proposed experiments involved several DGM trainings as well as the generation of large ensembles. Given the reasonable training and inference times involved in training GANs, GANs were chosen only to demonstrate the use of the proposed method. Recall, a GAN is a DGM that involves adversarial training between a generator network that creates new images, and a discriminator network that tries to distinguish the generated images from the training data.

It is anticipated that other researchers will employ the datasets made available with this work towards benchmarking other emerging DGMs as well as for studying the effects of various training strategies.

Two popular GAN architectures: ProGAN (PG) [119] and StyleGAN2: config-e (SG) [117] were employed for this work. The prescribed default training schedule was found to be sufficient for training in terms of visual quality, Fréchet Inception Distance (FID) 10k scores [45] and loss curve convergence. Recall that FID is currently the most popular measure of assessing DGM image quality. The trainings were performed such that the discriminator was shown 12 million images and 25 million images for PG and SG respectively; these were also the prescribed default training durations. For SG, the regularization parameter R_1 was set to 100 and the truncation parameter ψ was set to 0.5; both are default values for the chosen configuration. The trainings were performed on Nvidia GeForce GTX 1080Ti, 1080, Tesla V100 and A100 GPUs, and typically took between 2 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 SCMs for assessing common DGMs that are trained in a typical way.

3.4 Results

Sample generated images from both networks and all three SCMs are shown in Figure 3.4 while examples of artifacts are shown in Figure 3.5. The FID scores [45] for all models from both networks were between 2 and 10, indicating excellent image quality. (Lower FID scores are better and 0 indicates perfect image quality, state-of-the-art DGMs typically achieve single digit FID scores on natural image datasets.) Ensemble intensity distributions were also well replicated in all generated (DGM-generated) ensembles.

3.4.1 Results from the Flags SCM

Post-hoc processing of the DGM-generated ensembles demonstrated that perfect match with the foreground templates was achieved for about 98% realizations, while occasional malformations via blending of foreground templates was observed in the remaining cases. However, the forbiddance rule in Equation 3.4 was always respected. 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. Here, the exclusion was automated and not manual. Specifically, each binarized realization was compared against all binarized class templates and the absolute error was computed. Only perfect realizations, that is, realizations with zero error with respect to exactly one class template, were retained. By excluding ill-formed realizations from subsequent analysis, we avoided conflating the effects of foreground malformations with other per-image, statistical errors.

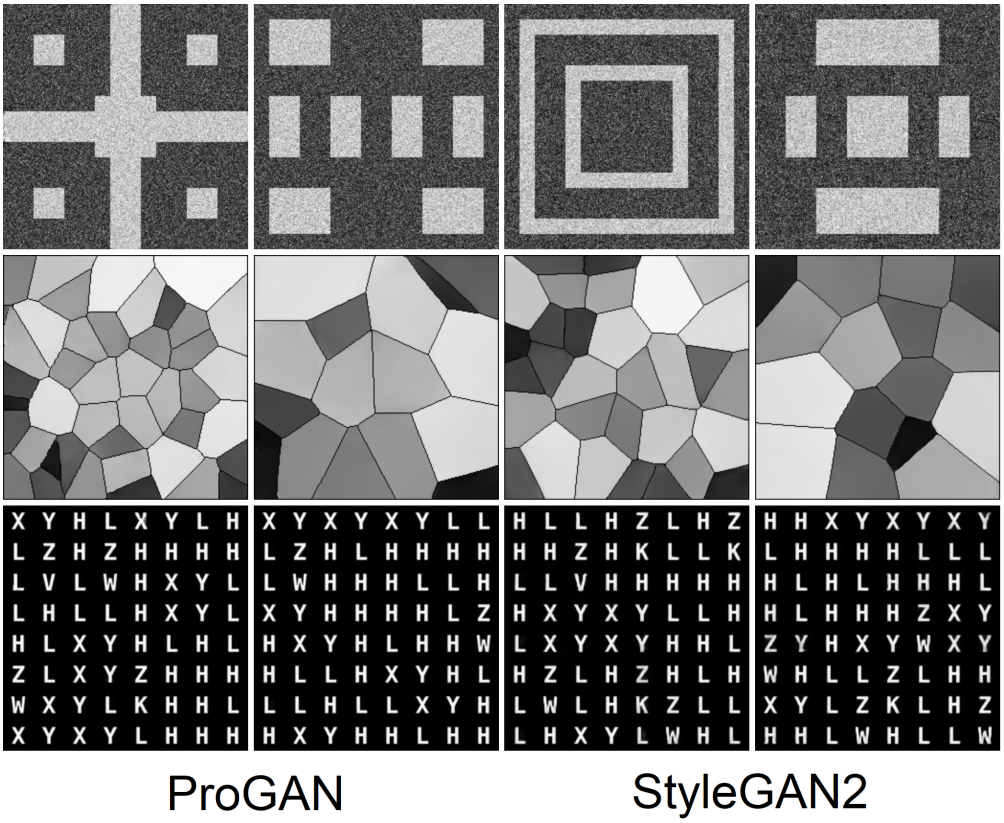


Figure 3.4: Subjectively visually good DGM-generated examples from networks trained on the three SCMs. Columns 1 and 2 show PG images while Columns 3 and 4 show SG images. Although the images demonstrate good visual similarity, contextual errors can be present in any image in any ensemble.

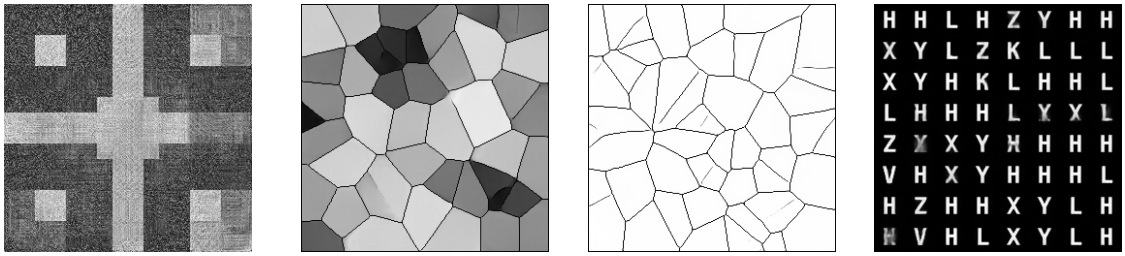


Figure 3.5: Class-mixing and artifacts in DGM-generated images. DGM-generated images occasionally exhibit artifacts such as blending of class-specific foregrounds in the flags SCM (left), weak boundaries and shading variance within distinct Voronoi regions (middle), and badly formed letters in the alphabet SCM (right).

Equation 3.2 and Equation 3.3, representing intensity distribution requirements, were tested against a generous tolerance of 99.5th percentile of the chi-square statistic computed separately for f and b. None of the realizations generated from either network satisfied Equation 3.2, i.e., the prescribed foreground distribution, while about 1% and 91% images violated Equation 3.3, i.e., the prescribed background distribution for PG and SG respectively, suggesting that the foreground and background were learned differently. This further indicates that first-order statistics 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 feature-specific foreground and background intensity distributions 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 prescribed randomness in pixel placement, was tested via the tile-wise computation of Moran's I (MI) of spatial autocorrelation [168] for each f_k and b_k in every I. A *tile* was considered acceptable if the MI was within $0 \pm \sigma_M/256$, where σ_M is the standard deviation of the distribution of the MI computed on the training data, and a *realization* was considered acceptable if at most 3 tiles were rejected. On average, 3% and 11% of the realizations violated the distribution of MI for the foreground and background for PG, while the proportion was about 4% for both subsets for SG. 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 9, of the realizations did not exhibit the prescribed randomness, and thus, inference based on the presumption of randomness could be incorrect. It was observed that the mean class prevalence matched the expected mean of 1/8, corresponding to uniform class prevalence in the training ensemble. Although the standard deviation was likely negligible for PG (σ=1%), it was non-negligible for SG (σ=9%), indicating that some classes were preferentially generated in the latter case. Thus, the relevant prevalence in a training data might not be reproduced in a DGM-generated ensemble—this might have significant implications when employed for data augmentation or statistical power calculations. Thus, *for this one SCM*, both second-order features and the per-image prevalence of second-order features were reasonably well reproduced; however, the first-order information per-image was essentially always wrong, even though the ensemble mean intensity distribution appears correct.

3.4.2 Results from the Voronoi SCM

Although high visual similarity was observed in the DGM-generated Voronoi images, various artifacts were also observed such as: the presence of (i) low-amplitude, high-frequency artifacts in regions of constant intensity [169], (ii) curved or "floating" region edges and (iii) multiple intensity values instead of only one in a single Voronoi region (see Figure 3.5).

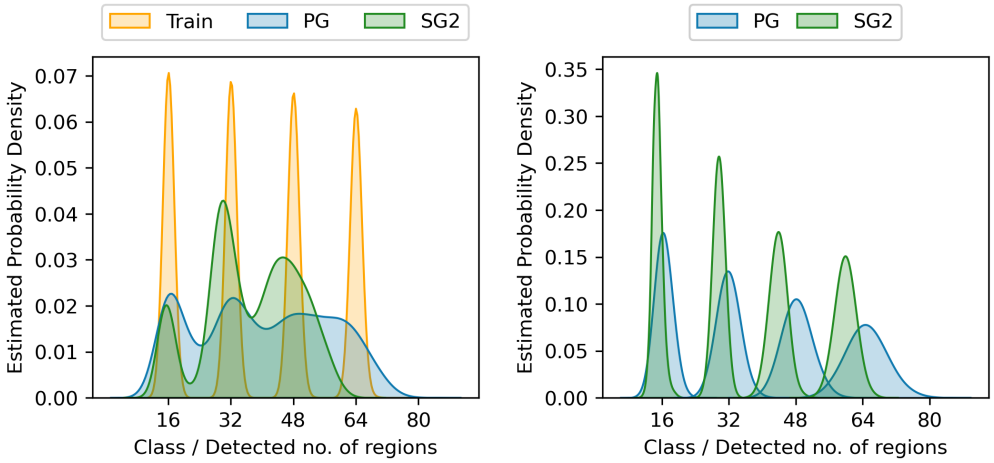


Figure 3.6: Results from the Voronoi SCM for class prevalence studies. Note that both sub-figures represent kernel density estimates of the corresponding distributions and hence the modes in the data appear Gaussian. Left: The expected equal class prevalence in the ensemble was not reproduced in the DGM-generated images from both networks, but more significantly for SG. The effect of error from the post-hoc classifier is also observed. Right: Four separate models, each trained on a single class, generated images outside the class for both networks. While the SG-generated ensemble demonstrated class extrapolation, the PG-generated ensemble showed slightly shifted class means.

Low-amplitude, high-frequency artifacts, possibly characteristic of the convolutional network architecture, could affect decision-making. This is because the presence of high-frequency artifacts impacts local statistics to some extent, that is, the original second-order information—and thus, possibly, texture statistics—may not be consistent with the original dataset. The other, more visually apparent artifacts, might confound a variety of classifiers or analyses which are calibrated on the training data. The high-order rule in Equation 3.5 relating intensity and area of a shaded Voronoi region also was tested; it was observed that the rule was not reproduced exactly. The expected Spearman rank correlation (ρ=1.0) was lower in the DGM-generated images. A decrease of over 20% ($\rho < 0.8$) was observed in 3% and 2% realizations from PG and SG, respectively. If the grayscale intensity g represents

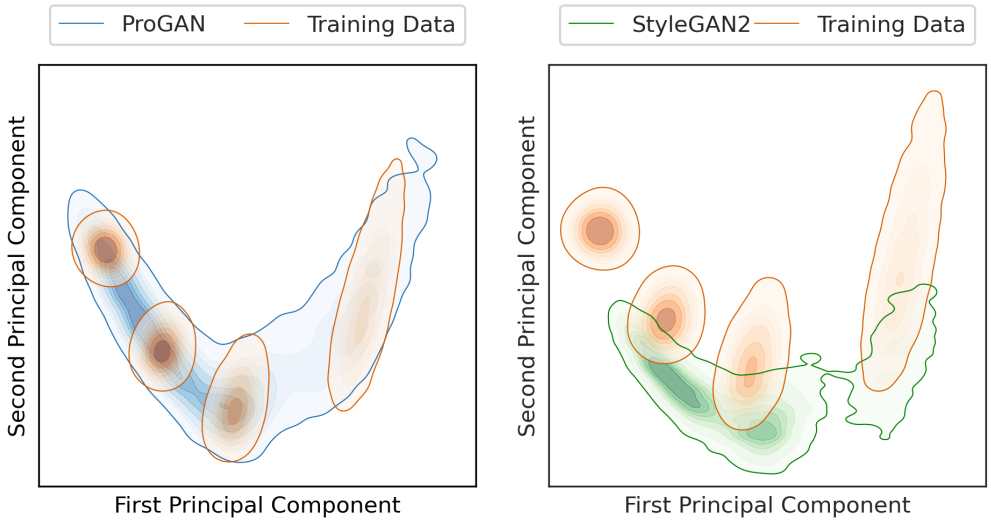


Figure 3.7: Results from the Voronoi SCM for assessment of implicit context. Statistics representing implicit context were projected onto the two highest principal components for true and PG-generated (left) and SG-generated (right) ensembles. Interpolation between the four training classes was more prevalent in the PG-generated ensemble while lower overlap in feature clouds was observed in the SG-generated ensemble, indicating dissimilar ranges of these statistics in the latter.

a physical property, violation of Equation 3.5 implies that these realizations have at least partially lost their quantitative meaning. This result might have serious implications for quantitative imaging modalities such as CT and PET, and is discussed later in this chapter subsection 3.4.4. Next, studies of class prevalence were performed with the Voronoi SCM by training five different models for each network architecture on the training data representing: (i) all four classes equally, and (ii-v) each of the four classes individually. As seen in Figure 3.6, class prevalence in case (i) was not maintained in the ensemble generated from either network (Figure 3.6 left) whereas class extrapolation was observed in cases (ii-v) (Figure 3.6 right).

Last, implicit context in Voronoi diagrams was assessed for case (i), i.e., DGMs trained on a four-class dataset with uniform prevalence. Several mathematical properties arise implicitly in Voronoi diagrams. Voronoi is a unique solution to a space partitioning problem and its properties are well-established. This implies that in order for an image to be a Voronoi diagram, certain statistics must co-occur in a specific manner. In other words, the implicit context must be correct. Two studies were undertaken to assess the reproducibility of implicit context. In the first study, several statistics were extracted and their conditional

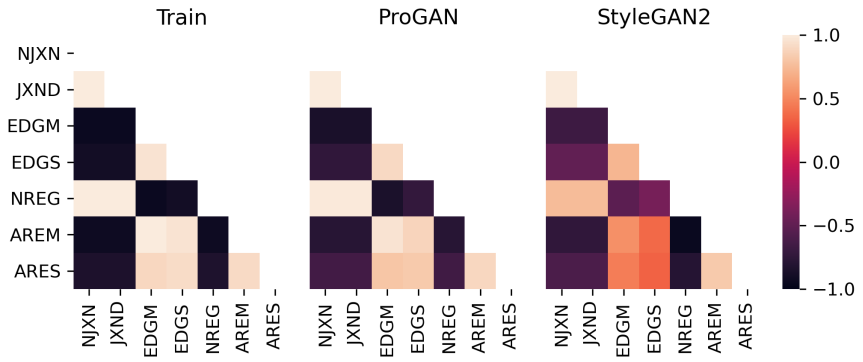


Figure 3.8: Results from the *unshaded* Voronoi SCM for assessment of implicit context. The strengths of correlations of the per-image statistics representing implicit context (left) were lowered in both DGM-generated ensembles (center and right), especially in SG, indicating that the correct implicit context was not reproduced.

co-occurrences were studied. In the second study, known mathematical properties of Voronoi diagrams were tested.

In the first study of conditional co-occurrences of Voronoi statistics, the Skan Python library [170] was employed to compute the following statistics derived from Voronoi regions and edges: number of junctions (NJXN), junction density (JXND), mean edge length (EDGM), standard deviation over edge lengths (EDGS), number of regions (NREG), mean area of a region (AREM), and standard deviation over region area (ARES). These statistics were chosen because they can be employed to study certain established properties of Voronoi diagrams. Interpolation between class-specific features was observed (see Figure 3.7) via principal component analysis (PCA) of the features listed above. This indicates extrapolation in the feature space corresponding to implicit statistics learned by the DGMs. Even when the classes (or NREG) were incorrect due to extrapolation, the implicit context was generally retained in this case.

Reproduction of implicit context was then tested in the absence of shading. Partial loss of implicit context was observed via decreased correlations between the studied statistics (see Figure 3.8) and lower overlap in the feature clouds in PCA (not shown) as compared to Figure 3.7. In Figure 3.8, the training data (left) demonstrates nearly perfect cross-correlation or anti-correlation between all pairs of studied statistics, indicating that if the value of one statistic is changed, the values of all other statistics would be required to change accordingly to maintain the properties of a true Voronoi diagram. However, the strength of

correlations is not maintained in case of both DGMs, but particularly StyleGAN2 (right). This indicates that there are certainly images in the SG2-generated ensemble that cannot be considered Voronoi diagrams.

Further confirmation of implicit contextual errors in the DGM-generated ensemble was obtained via testing two well established statistical properties of Voronoi diagrams in the second study of implicit context. Specifically, Property V11-1 and V12 presented in Bootes et al. [167] (here onward referred to as P1 and P2) were tested. These properties respectively are:

$$n_e \leq 3n - 6, \tag{3.9}$$

and

$$n_v \geq \frac{1}{2}(n - n_c) + 1, \tag{3.10}$$

where n_e, n_v, n_c, and n, indicate the number of edges, the number of vertices, the number of bounded Voronoi regions, and the number of all Voronoi regions respectively within a given image. Both conditions were satisfied in 99% of the training data. For the unshaded Voronoi, the rates of violation for both conditions in PG-generated and SG-generated ensembles respectively were: 12% and 100% for P1, and 10% and 100% for P2. For the shaded Voronoi, these rates were under 6% for both DGM-generated ensembles. Note that it is possible that a different model or training strategy may have fewer implicit contextual errors. Here, it is only demonstrated that implicit contextual errors made by a model trained in a typical manner and achieving low FID scores, can be detected via the proposed method. These results suggest that the reproduction of implicit context even for datasets such as the unshaded Voronoi is a non-trivial task for a DGM and may have significant implications in domain-specific space partitioning problems.

3.4.3 Results from the Alphabet SCM

Although most letters were well-formed in the DGM-generated ensembles, errors occasionally were observed as seen in Figure 3.5. Error rates of letter formation via post-hoc processing were: 1 in 6250 letters for PG and 1 in 73 letters for SG respectively, indicating that almost all letters in a realization were recognizable. Although relatively few letters were unrecognizable,

only images where all letters were recognizable, 99% and 59% of the ensemble for PG and SG respectively, were considered for further analysis. A sample of 10000 such well-formed realizations was explicitly tested for high-order rules of feature prevalence.

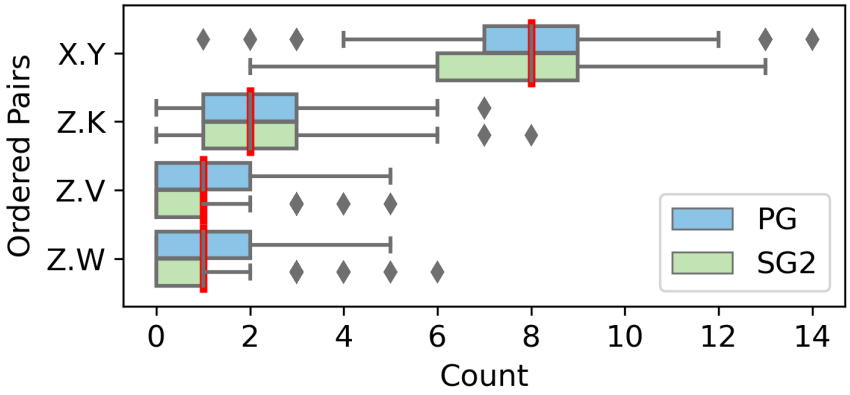


Figure 3.9: Results from the alphabet SCM. The expected paired-letter prevalences: X-Y, Z-K, Z-V, Z-W = 8, 2, 1, 1 were not respected by either network. Correct prevalence is marked in red. A wide range of values was seen for both networks indicating that perfect prevalence is achieved only by chance.

The observed frequency of letters was compared to the prescribed set $\mathbb{B}$ via the χ^2 goodness-of-fit test. Only 119 PG and 72 SG 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, on testing per-image letter prevalences (Equation 3.6), it was observed that only 18 PG realizations and 6 SG realizations exactly matched $\mathbb{B}$; recall, this frequency is identical in *every* training image. Thus, only by rare chance was any realization correct in high-order. Incidentally, it was observed that these were not memorized realizations. In a certain domain, if natural variation exists in the prevalence of a feature, most realizations would be acceptable. However, if the feature prevalence is the context required for a downstream task, then essentially none of the realizations from these DGM-generated ensembles are acceptable. Next, the prescribed ordered pair-prevalences (Eqs. 8 and 9) were tested. The fixed ordered pairs X-Y, Z-K, Z-V and Z-W were expected to occur at frequencies of precisely 8, 2, 1 and 1 respectively, but were observed to occur at a wide range of frequencies (see Figure 3.9) for both networks. The single letters V, W and K which never occur without the partner in the training data, occurred without the other member of the pair up to 100% of time. Similarly, the letter Y occurred separately about 37% of time for PG. This rate of separate occurrence of letters in letter-pairs is approximately doubled or tripled for SG. Hence, pairs of image features

that may be expected to have known, relative locations and prevalence might not appear in a DGM-generated ensemble. Thus, "visually good" DGM-generated images might have diminished domain-specific value due to an unrealistic representation.

3.4.4 Interpretation of Results within Biomedical Imaging

In biomedical 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, the joint reproduction of statistics across multiple orders might be a challenging task for the chosen DGMs as observed in the results from the flags SCM 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, a DGM employed for simulating positron emission tomography images of a certain tumor type may produce a majority of tumors of correct shape but significantly different in the expected intensity distribution and texture. Drawing diagnostic inferences from such a generated ensemble, even when employed for data augmentation, might translate to false clinical predictions.

The Voronoi SCM 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. When the rank-correlation between area and grayscale intensity was not reproduced correctly, the quantitative information—here, representative of physical tissue properties—could be unreliable and the derived textural features suspect. Furthermore, for both multi-class SCMs, flags and Voronoi, the incorrectly reproduced class prevalence in the generated ensembles suggests that if these particular instances of DGMs were used to replicate a clinical dataset for virtual clinical trials, or to generate a training ensemble for a downstream task, 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. It is hypothesized that the DGM loss function as well as the dimensionality of the latent space may contribute to the learning of some image statistics but not others. However, it is emphasized that the relation between a network architecture and per-image statistics is an open research question. The method proposed in this chapter can potentially expose the lack of capacity of a network architecture for learning certain orders of image statistics. The proposed datasets

can aid in improving network design by serving as test datasets while prototyping novel and task-specific architectures.

The alphabet SCM was designed with known per-realization prevalence of single and paired features in order to isolate reproducibility of high-order features (the letters) from that effects of variable position, structure and shading. Recall the earlier example of a wrist radiograph. Just as the Capitate and Hamate bones always articulate with each other, the letters X-Y also are always expected to occur only as horizontal, ordered pairs. 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 as was observed in the DGM-generated images of the alphabet SCM where less than 0.2% of the ensemble had perfect feature and feature-pair prevalence. Such visually realistic DGM-generated images with incorrect per-realization feature prevalence might have reduced diagnostic value. This could translate into a bias in, or even a complete failure to learn, the information required for particular decision tasks.

3.5 A Brief Exploration of Stochastic Context Models for the Evaluation of Image-Conditioned DGMs

One solution to improve contextual correctness in DGM-generated images is to condition the generation on relevant domain knowledge via an image. The rationale is that, ideally, if sufficient domain-relevant context [171] is made available via the input image employed for conditioning, an image-conditioned DGM (IC-DGM) [144, 172] could potentially learn to generate contextually correct images. In this section, the idea of assessing contextual reproducibility via SCMs is extended to image-conditioned DGMs (IC-DGMs) for domain transfer tasks with the goal of quantifying domain-relevant contextual errors that persist despite image conditioning. A second motivation for this short study is that similar to the evaluation of unconditional DGMs, the evaluation of IC-DGMs remains an open challenge [173,174], despite their popularity in biomedical imaging research [1]. An SCM-based method of evaluation could provide insights into the capacities of IC-DGMs.

3.5.1 Methods

Two SCMs were adapted to assess image-conditioned DGMs (IC-DGMs), namely, the Voronoi SCM (V-SCM) and the Alphabet SCM (A-SCM).

Recall that the V-SCM was designed to encode per-image context via pre-specified rules of shading, thus lending grayscale values contextual quantitative significance. This is analogous to histology images, where a one-to-one correspondence may exist between cell types, their sizes, and stains. For assessing IC-DGMs, the shaded Voronoi SCM was paired with an unshaded version of the same SCM, which was employed for conditioning. Thus, the implicit contextual information was already provided to a DGM via conditioning.

The adapted A-SCM, here onwards referred to as A-SCM2, is designed to represent context explicitly as per-image feature prevalences represented under various inter-domain mappings (e.g., bijective, surjective), without complex shading variations. A more complex version of this SCM would be analogous to transforming images across imaging modalities. For example, when the task is to generate PET images given CT, there are several kinds of mappings present for different organs in the two imaging modalities. Some organs might be visible reliably in both imaging modalities, and hence might potentially appear correct in shape and size in the generated PET images. Other organs may be present only in CT but not in PET; here, a CT to PET transform might be potentially accurate as before, but a PET to CT reverse transformation might be infeasible. Furthermore, as PET represents functional activity (unlike CT), the intensity variations related to organ function, but independent of anatomy, might be inaccurate. The A-SCM2 is aimed to represent logically similar unique and non-unique mappings between structures in two imaging domains to test the contextual correctness in the generated images.

All images generated from the SCMs described below were size 256×256 pixels, 8-bit grayscale images.

Description of the Voronoi SCM for Domain Transfer Tasks (V-SCM2)

The four-class V-SCM2 for domain transfer consists of matched image-pairs of unshaded and shaded Voronoi diagrams [167]. The shaded Voronoi is the same as described in section 3.3.1. For the unshaded Voronoi, the grayscale values were set to 255 for the background, and 0

for the Voronoi region boundaries. The task for a IC-DGM then, was to generate a correctly shaded image given the unshaded Voronoi image.

Description of the Alphabet SCM for Domain Transfer Tasks (A-SCM2)

A realization described by the A-SCM2 consisted of 32×32 pixel tiles t arranged in a regular 8×8 grid. Each t represents a letter from a pre-specified set: $\mathbb{A} = \{H, K, L, V, W, X, Y, Z, \square\}$, where the last element represents a blank tile, which enables the representation of the absence of any information as a feature, similar to blank spaces observed in many biomedical images. Furthermore, exact per-image letter and letter-pair prevalences were specified to constitute a domain $\mathbb{D}$. Generally, any realization from a domain $\mathbb{D}$ can be described as:

$$I = \{t_{r,c} : \bigcup_{r,c} m(t_{r,c}) = \mathbb{D}\}, \tag{3.11}$$

where $m(t) : t^{32 \times 32} \rightarrow \mathbb{A}$ describes template matching, which was also employed for post-hoc analysis, and r, c are respectively the row and column indices of t within the grid. Two such domains: $\mathbb{D}_1$ and $\mathbb{D}_2$, were specified in each experiment described below such that a forward mapping exists from $\mathbb{D}_1$ to $\mathbb{D}_2$. In all three experiments described below, the domain transfer task for a IC-DGM was to recover the *inverse* mapping and generate an image: I_{out} in $\mathbb{D}_1$, given another image: I_{in} in $\mathbb{D}_2$; the differences in experiments lie in the contextual rules of per-image letter prevalences, placements, and inter-domain mappings (bijective or surjective). Samples from the A-SCM are shown in Figure 3.10.

Experiment 1 (E1): $\mathbb{D}_1 \leftrightarrow \mathbb{D}_2$

The goal of E1 was to test the capacity of IC-DGMs to generate a contextually correct image in $\mathbb{D}_1$ when (i) inter-domain letter-pair mappings are bijective, (ii) single letter mappings from $\mathbb{D}_1$ to $\mathbb{D}_2$ are surjective, and (iii) sufficient context was included in I_{in} from $\mathbb{D}_2$ for unique recovery. Realizations in $\mathbb{D}_1$ were generated to have the following fixed per-image prevalences: $\mathbb{D}_1 = \{8 \times \{XY\}, 2 \times \{ZK\}, 1 \times \{ZV\}, 1 \times \{ZW\}, 16 \times L, 24 \times \square\}$. The adjacency rules for the ordered letter-pairs in $\mathbb{D}_1$ were:

$$m(t_{r,c}) = X \Leftrightarrow m(t_{r,c+1}) = Y, \tag{3.12}$$

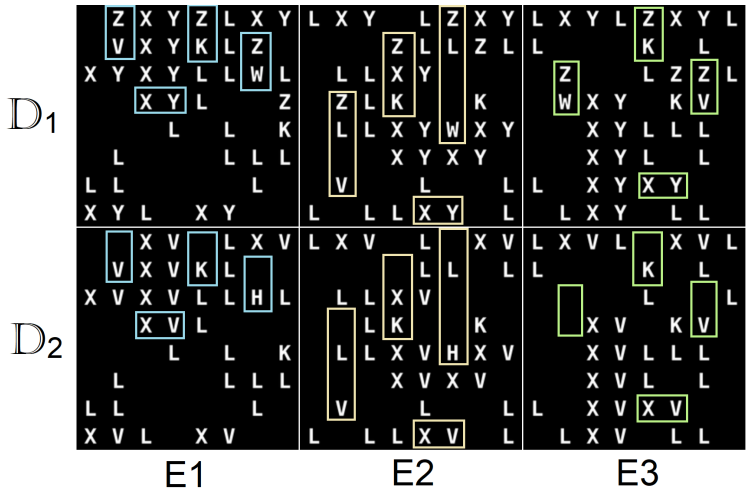


Figure 3.10: Samples from experiments employing the A-SCM2. Experiments E1 and E2 map $\mathbb{D}_1$ to $\mathbb{D}_2$ bijectively, while E3 involves a surjective mapping. In all experiments, the forward mapping: $\mathbb{D}_1$ to $\mathbb{D}_2$ is prescribed in the training data, and the task for the IC-DGM is to recover the inverse mapping and generate an image in $\mathbb{D}_1$, given an image from $\mathbb{D}_2$. Note that the colored boxes highlight an instance of the letter-pair transforms in each experiment to aid the reader but are not actually present in the data.

$$m(t_{r,c}) \in \{V, W, K\} \Leftrightarrow m(t_{r-1,c}) = Z. \tag{3.13}$$

Next, for all images in $\mathbb{D}_1$ generated as described above, matched images in $\mathbb{D}_2$ were created via the following transformations:

$$\begin{aligned} XY &\mapsto XV, \\ ZV &\mapsto \square V, \\ ZW &\mapsto \square H, \\ ZK &\mapsto \square K. \end{aligned} \tag{3.14}$$

Thus, although the letter-pair mappings are unique between the two domains, the individual letters might not be uniquely transformed. For example, Y and V in $\mathbb{D}_1$ both map to V in $\mathbb{D}_2$. Yet, the unique recovery of I_{out} in $\mathbb{D}_1$, given I_{in} in $\mathbb{D}_2$ is expected when the spatial context of a letter is considered, which in this case is whether the left neighbor of the letter V in $\mathbb{D}_2$ is the letter X.

In summary, the DGM is asked to transform individual letters from one domain to another, based on a unique mapping between the two domains. To learn the unique mapping, the DGM has to learn not just mappings between individual letters across the two domains, but also account for their neighborhoods, or contexts, which determine the transformation. Because a unique mapping exists between the two domains, this experiment serves as a first test for any DGM that is considered for domain transfer. If the DGM cannot perform this simple domain transfer task, it may not have the capacity for more complex tasks with non-unique solutions. In real-world tasks, bijective mapping is rarely present between two imaging modalities.

Experiment 2 (E2): $\mathbb{D}_1 \leftrightarrow \mathbb{D}_2$

The goal of E2 was to assess whether the expected, unique image recovery in E1 can still be achieved by a IC-DGM when the relevant spatial context manifests at larger length-scales as spatially distant letter-pairs. To generate realizations in $\mathbb{D}_1$, Equation 3.12, Equation 3.13, and Equation 3.14 in E1 were replicated but the specified vertical pairings in Equation 3.13 were modified as:

$$\begin{aligned} m(t_{r,c}) = K &\implies m(t_{r-2,c}) = Z, \\ m(t_{r,c}) = V &\implies m(t_{r-3,c}) = Z, \\ m(t_{r,c}) = W &\implies m(t_{r-4,c}) = Z. \end{aligned} \tag{3.15}$$

Thus, the pairs ZK, ZV and ZW were separated by 1, 2 and 3 letter-tiles respectively. Matched images in $\mathbb{D}_2$ were obtained via transformations described in E1.

In summary, the second experiment also assesses a DGM for domain transfer under a simple, bijective (i.e., unique) mapping between domains. The individual letter transformation can be learned from the neighborhood of individual letters as in E1, but the neighborhoods in E2 can be upto four times larger than in E1. This enables the testing of network capacity for capturing very large features or conditionally co-occurring features that are distant from each other. If the network truly learns all conditional co-occurrences within each image, it should demonstrate perfect performance on this task. Note that the unique transformation can be recovered via logical rules, and does not require a DGM. The knowledge of logical rules, and hence, the "right answer" enables the *testing* of a DGM.

Both experiments, E1 and E2, are logically similar to a domain transfer task where the first domain contains sufficient information to accurately generate an image in the second domain, and thus, a DGM should achieve perfect performance in both these experiments.

Experiment 3 (E3): $\mathbb{D}_1 \twoheadrightarrow \mathbb{D}_2$

As compared to E1 and E2, E3 explores domain transfer within a slightly more realistic inter-domain mapping: a surjective contextual mapping, represented via letter-pairs (in addition to single letter mappings being surjective like E1 and E2). Thus, the IC-DGM has a purposely unfair domain transfer task in terms of learning to *place* a certain letter-pair. However, because the per-image *prevalences* of letter and letter-pairs in $\mathbb{D}_1$ were fixed, ideally, they should be correctly reproduced in the generated ensemble even if the locations of the expected letter-pairs were incorrect. Specifically, the only rule in E3 that differed from E1 was the transformation: $ZW \mapsto \square\square$, which resulted in a non-unique mapping between the two domains. This study is logically similar to the unfair domain transfer task of generating CT images given PET images, i.e, when the latter domain does not contain sufficient information to accurately generate an image in the former domain.

Although this experiment studies more realistic mappings than E1 and E2, the domain transfer task in E3 does not have a unique solution. This task is unfair in terms of transformation of entire images, but, (i) some features should be reliably transformed, and (ii) features expected in the output domain should be correct in terms of their prevalences. Thus, this experiment is an important test which exposes the kinds of mistakes that may be expected from DGMs employed for domain transfer in realistic scenarios.

Network Trainings

Two popular IC-DGMs: Pix2pix [175] and CycleGAN [176] were employed with V-SCM2 and A-SCM2 ensembles sized 65536 and 131072 respectively, split into training (75%), validation (12.5%) and test (12.5%) sets. All default training choices were retained, unless specified in the experiments. Data pre-processing and augmentation during training was disabled. Both networks were trained for 10 epochs in all experiments except A-SCM2: E3, in which the networks were trained for 15 epochs. A model was chosen for inference based on the

highest visual similarity of the generated images. All trainings were performed on Nvidia A100 GPUs and took between 3 to 12 hours on a single GPU.

3.5.2 Results from the Adapted Voronoi SCM: V-SCM2

Samples generated from the two IC-DGMs are shown in Figure 3.12. Visually, the shading rules appear to be well captured. However, some undesirable effects such as shading gradients within a single Voronoi region, or low contrast in some image areas due to weaker area-grayscale correlation, were observed. Checkerboard artifacts, known to occur with the original CycleGAN architecture, were also observed; their effects on the shading task were included in the quantitative results.

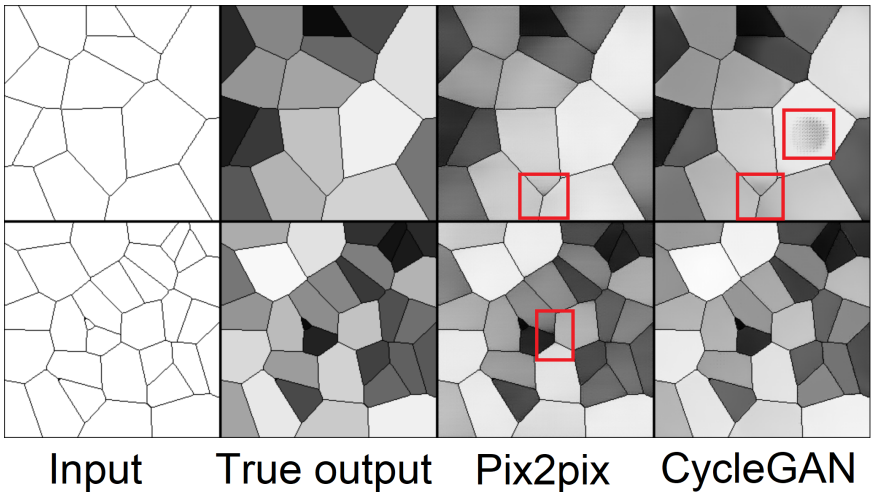


Figure 3.11: Sample images from the two DGM-generated ensembles of V-SCM2 are shown. Artifacts such as checkerboard, shading gradients, and low contrast are highlighted.

First, we tested the joint replication of the two pre-specified contextual shading rules: perfect correlation of grayscale with area in each image ($\rho = 1$) and no variation in grayscale value within a Voronoi region, i.e., grayscale standard deviation $\sigma \approx 0$. For both IC-DGMs, a wide range of values was observed for both rules, irrespective of the weight (λ) of the L_1 loss in Pix2pix or the identity loss in CycleGAN. However, in Pix2pix, increasing λ did appear to decrease the range of absolute error in ρ from almost 20% to 5% (see Figure 3.12, row 1) unlike CycleGAN, where this range was about 15% for all λ. In the training data, this range is exactly 0 and $\rho = 1$ in all images.

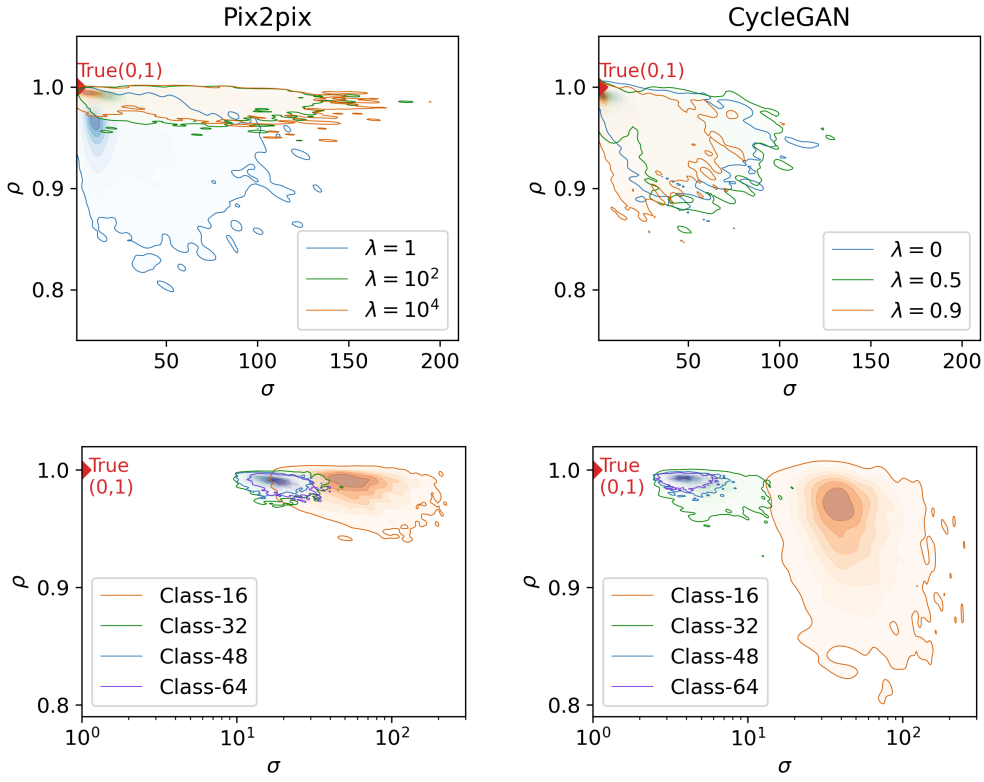


Figure 3.12: Results from the V-SCM2. Row 1: Neither network exactly replicated the grayscale correlation with area ($\rho = 1$) and the constant grayscale intensity in each region ($\sigma = 0$) for varying loss function weights. Row 2: Class-wise analysis indicates that the errors are more widespread for class 16 as compared to the higher classes. Note the log-scale for X axis to highlight class differences. The true value from training data is marked in red.

These results demonstrate that the quantitative value of the DGM-generated images may be partially lost and the extent of this loss can be explicitly quantified before a IC-DGM is considered for domain-specific deployment. Next, we assessed the class-wise performance for the intermediate (also default) values of λ in the previous experiment. For both networks, class 16 demonstrated remarkably greater variations in both shading properties than the other classes (see Figure 3.12, row 2). Particularly, class 16 demonstrates mean σ values that are 4 times (Pix2pix) and 15 times (CycleGAN) worse as compared to class 64. Thus, via class-based analysis, a correspondence between various length-scales and contextual shading accuracy can be established for a IC-DGM. In this way, the V-SCM2 can further enable an objective choice of an architecture in a human interpretable manner before IC-DGM-deployment, especially if the domain-relevant length-scale is known.

3.5.3 Results from the Adapted Alphabet SCM: A-SCM2

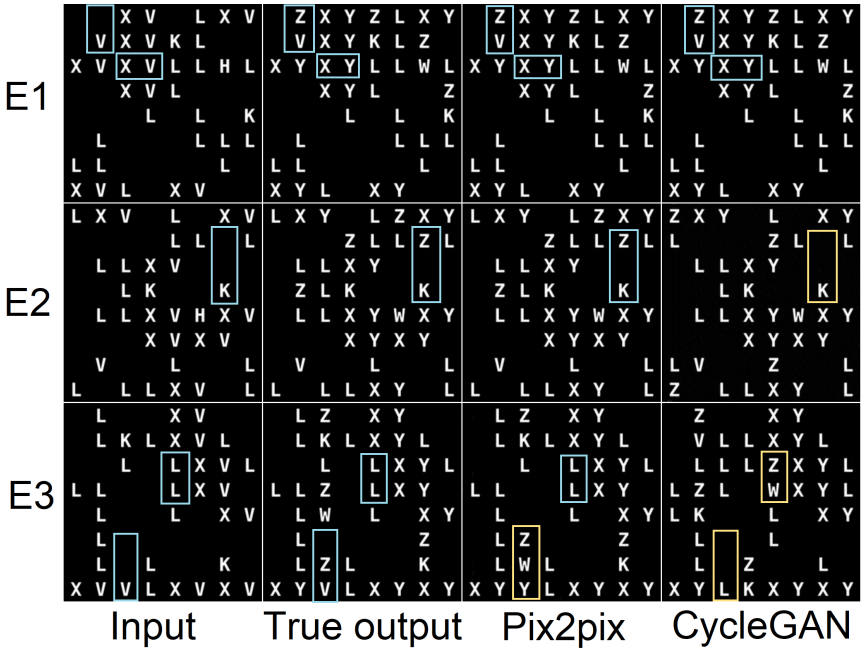


Figure 3.13: Results from the A-SCM2. Although the DGM-generated images show high visual similarity, contextual errors are present in E2 and E3. Some examples of perfect letter-pair recovery are highlighted in blue while errors are highlighted in red.

Examples from the training and DGM-generated ensembles are shown in Figure 3.13. Although the letters appeared well-formed in all cases, some contextual errors were still observed. Results from all experiments are reported in Table 3.1. Perfect recovery of all letter-pairs by both networks was observed in E1, where $\mathbb{D}_1 \leftrightarrow \mathbb{D}_2$, and letter-pairs were adjacent, i.e., present within the receptive field of the default PatchGAN discriminator. However, in E2, when these letter-pairs were spatially distant, only Pix2pix demonstrated perfect recovery of all letter-pairs, including a letter-pair (ZW) that was separated by 128 pixels; CycleGAN failed to recover this most distant letter-pair in *all* DGM-generated realizations. These results imply that the presence of an inter-domain bijective mapping alone is insufficient for accurate domain-transfer, the capacity of the IC-DGM to capture the relevant context present in the input image is also essential [177]. In E3, errors were expected to occur in the placement of the surjective mapped letter-pair (ZW). By chance, only about 2% of the Pix2pix-generated realizations, and none of the CycleGAN-generated realizations demonstrated perfect recovery in both: placement and prevalence. Besides the placement errors of this letter-pair (ZW), errors were also observed in its per-image prevalence (see Table 3.1). Furthermore, contextual errors were also observed in the prevalence of some bijectively mapped letter-pairs, constituting up to 25% (Pix2pix) and 70% (CycleGAN) of the DGM-generated ensemble. The issue of recovering bijectively mapped letter-pairs was alleviated entirely (Pix2pix) or partially (CycleGAN) by replacing the PatchGAN discriminator with an ImageGAN, which can access the entire image in its receptive field as opposed to image patches in PatchGAN. However, the ImageGAN discriminator did not improve the per-image prevalences of the surjective mapping. Thus, for a IC-DGM, the A-SCM2 enables not only the quantification of error-rates of per-image feature prevalences for various inter-domain mappings and spatial extents of context, but may also enable the quantification of features that are correct by chance, before it is consideration for practical deployment.

3.5.4 Interpretation of Results

In this brief study, it is not claimed that the results are generally applicable to the chosen IC-DGMs in all cases, but that they apply to an instance of a reasonably trained model chosen to demonstrate the method of evaluation. Furthermore, the experiments were designed to demonstrate the capabilities of the SCMs towards the evaluation of IC-DGMs; statistically rigorous and systematic studies employing these SCMs will be undertaken in future.

Table 3.1: Results from the A-SCM2. Accuracy (%) of the per-image letter-pair prevalence within the the DGM-generated ensemble in each experiment employing the default IC-DGM architecture is reported; E1: all pairs, E2: most distant letter-pair, E3: (i) best reproduced, (ii) worst reproduced bijectively mapped letter-pair, and (iii) the surjectively mapped letter-pair. Results with an ImageGAN discriminator (IG) are also reported for the last case (E3-iii). Bijective and surjective mappings are indicated by "bij" and "sur" respectively.

Expt.	Pair mapping	Context (pixels^2)	Pix2pix (% acc.)	CycleGAN (% acc.)
E1	bij	32×32	100	100
E2	bij	128×32	100	0
E3	bij	32×32	(best) 100	100
	bij	32×32	(worst) 74	30
	sur	32×32	41	69
E3(IG)	sur	32×32	44	31

Specifically, the V-SCM2 enables the quantification of contextual errors in shading and also relates them to specific length-scales, thus providing a first test before employing IC-DGMs on domain-specific tasks that involve quantitatively significant shading, e.g., generating histology images with cell-specific shadings [178]. The A-SCM2 provides a method to assess the reproducibility of per-image contextual features and their prevalences in DGM-generated ensembles under a variety of inter-domain mapping conditions and context sizes. Such an evaluation of network capacity may prove beneficial to rule out DGMs for domain-transfer tasks that require significant domain expertise for visual evaluation. In general, results from both DGMs suggest that image-conditioning alone may not necessarily alleviate contextual inaccuracies in the generated image ensembles.

For a domain-transfer task, a domain-specific interpretation of results from the SCM-based method of evaluation may be obtained if the following are identified: (i) the length-scale of relevant spatial context in the domain of deployment, (ii) inter-domain mapping type (e.g., surjective), and (iii) tolerance of error-rates in the generated ensemble. In this way, the SCM-based method may aid the objective choice of a network for a given domain-specific task.

3.6 Discussion

Much improvement has occurred in the realism of DGM-generated natural images and their evaluation [179,180]. However, the deployment of DGMs in domains where domain expertise is inextricably tied to image perception, such as in medical imaging, still remains a challenge [181]. To partially circumvent this challenge, the proposed SCMs provide a method for encoding high-order information relevant to a domain while also allowing the recovery of this information from a DGM-generated ensemble. In other words, the proposed SCMs provide a kind of domain-relevant "ground truth" for assessing DGMs. In the present work, high-order information was represented via explicit modeling of contexts such as feature prevalences and relative feature arrangements, but the use of SCMs is not limited to these scenarios. The proposed method is general in the sense that any other representation of spatial context that may be relevant to a certain domain could be employed similarly for evaluation as long its recovery from the generated ensemble is sufficiently robust. New SCMs, other than the three SCMs proposed in this work for unconditional image synthesis, can be designed by other researchers to encode any spatial context of interest or relevance in their domain. In addition, results from the three SCMs can be interpreted for applications in several domains. The SCM-based method of DGM evaluation can also be further developed for a specific domain as demonstrated in [137] for medical imaging.

Some works have studied the reproducibility of long-distance spatial context by generative models [122,182]. However, these methods do not purposefully design synthetic datasets for evaluation like those proposed in this work. One work [183] has assessed the generalizability of GANs for a few contextual attributes such as color and per-image prevalence in RGB images. Our proposed SCMs encode a wide range of contextual constraints at multiple orders of information. As the proposed method of evaluation is data-centric and independent of the generative model type, it can be readily employed on any generative model. Thus, the proposed method may enable the benchmarking of new architectures against existing architectures or aid the design and development process of generative models for domain-specific applications.

Of course, each instance of a particular DGM is unique, and thus the results may vary between instances, however, any instances trained from the same architecture share some common learning capacity. The use of designed SCMs is envisioned as a kind of "necessary but not sufficient" triage of DGM capacity. Our supposition is that if a particular DGM

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 domain-relevant image that comprises those fundamental properties. This is why SCMs such as the ones proposed can be relevant to estimating the probability that a DGM has made errors in domain-specific images. In future, we intend to extend the method of evaluating DGMs via SCMs to tasks other than unconditional synthesis, such as conditional synthesis and de-noising. Some promising results from an exploratory study were reported in section 3.5.

3.7 Conclusion for Chapter 3

The main conclusion is that SCMs can be designed to enable the quantification of certain impactful, per-realization errors, i.e., hallucinations, made by some popular DGM architectures 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 scenarios requiring a substantial domain expertise is that there usually is not a mathematically specified ground truth or expert labeling for each generated realization.

In this chapter, it is demonstrated how stochastic context models can be purposefully designed to include known high-order contextual information, analogous to domain-relevant external 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 SCMs were proposed and employed in the evaluation of two popular DGM architectures.

Across various training and model scenarios, it was found that the tested models failed to simultaneously reproduce all prescribed contextual features, at once, despite being well replicated in the ensemble, and despite obvious visual similarity between training and generated data. Specifically, numerous per-realization errors occur in: grayscale intensity distribution, spatial arrangement of those intensities, and, perhaps most impactful, in the frequency of pre-specified rates of feature occurrence. Here, it is not claimed that one architecture is better than the another, but that observable differences between the chosen instances of the all architectures can be exposed by the use of the proposed method.

The corollary is that the designed SCMs can serve as a kind of triage before even more sophisticated task-based measures of generated image quality are employed, or as benchmarking datasets for advancing generative model design.

The following chapter expands the idea of assessing spatial context to more complex and realistic stochastic models that describe anatomical constraints.

Chapter 4

Employing a Stochastic Object Model for the Evaluation of DGMs in Biomedical Imaging

"When we try to pick out anything by itself, we find it hitched to everything else in the Universe." - John Muir

4.1 Overview

Stochastic models of context, as described in the previous chapter, enable the ruling out of DGMs that lack the capacity to learn certain prescribed contextual attributes. SCMs do not encode anatomical constraints but represent logically analogous scenarios to anatomy. The reasonable next step in DGM evaluation is to test DGMs on more realistic medical images. As opposed to SCMs, contextual features in medical images cannot be explicitly described even by an expert. However, some anatomical structures can be described statistically, and context arises implicitly from the interactions of these stochastic anatomical structures. Furthermore, changes in certain anatomical features are strongly associated with changes in certain other anatomical features, and one set of features cannot be modified without causing a certain effect in another set of features. These associations or correlations are captured via implicit contextual features.

In this chapter, a stochastic *object* model (SOM) of the human female breast: Virtual Imaging Clinical Trial for Regulatory Evaluation (VICTRE) [133] is employed along with an evaluation framework based on image statistics related to object recognition. The dataset and the framework together were employed to conduct a public Grand Challenge: the Deep

Generative Modeling for Learning Medical Image Statistics Challenge, organized by our lab and hosted by the American Association of Physicists in Medicine (AAPM). In this Challenge, participants submitted trained DGMs for evaluation. Participants had access to the designed dataset and were evaluated according to the implicit contextual framework described later in this chapter. In general, it was observed that the overall ranking of the submissions according to our evaluation method (i) did not match the FID-based ranking, and (ii) differed with respect to individual feature families. Another important finding was that different DGMs demonstrated similar kinds of artifacts. Results from this chapter highlight the importance of domain-specific evaluation to further DGM design as well as deployment. The results also suggest that a DGM that is the best choice for one task may not necessarily be the best choice for another task.

The design and organization of the Challenge was a collaborative effort. My primary responsibilities lay in the design of the Challenge, and involved: (i) dataset design (subsection 4.4.2), (ii) design of the evaluation framework (subsection 4.4.3), and (iii) additional analyses of the submissions as described in subsection 4.5.3, subsection 4.5.4, and subsection 4.5.5. The organizational/ logistical aspects of the Challenge are not a contribution of this thesis, but are reported for completeness.

4.2 Introduction

In the medical imaging domain, SOMs enable the evaluation of imaging systems through virtual clinical trials, facilitating the assessment and optimization of medical imaging systems [48, 184]. Although virtual trials may not entirely replace physical clinical trials, they offer an important alternative that might complement or reduce the burden of physical clinical trials for the assessment of novel medical imaging technologies [135]. For conducting a virtual imaging trial, models are required for inputs to the imaging system, and for the imaging system itself, in addition to an image analysis/interpretation process. SOMs can satisfy the first requirement. Several SOMs have been proposed; these include models of a human female breast [185] as well as the entire human body [186]. Unlike stochastic models of context, which explicitly encode the context to be tested, SOMs at least partially describe the statistics of a specific object to be imaged [104]. As a result, SOMs could potentially appear more realistic with respect to the object being imaged. The level of modeling complexity and

realism in SOM design may be determined by the purpose of the SOM. The design of SOMs is primarily aimed at image acquisition experiments, and typically is not informed by post-hoc analyses methods. When assessing DGMs, post-hoc recovery of domain-relevant features from the generated images is essential for evaluation, which may not be guaranteed in SOMs. Hence, SCMs (as described in the previous chapter) were designed with the rationale that a *recoverable* ground truth could be encoded in a stochastic model to enable the assessment of DGM-generated images.

When complex SOMs are employed as training data for DGMs, there are three main challenges. First, a mathematical describable ground truth is not available. Second, statistics that completely represent an anatomy are not known, e.g., breast tissue cannot be exactly described via a set of known statistics. In such scenarios, assessing implicit contextual features as opposed to explicit contextual features provides a solution. Third, because typical SOMs do not take into consideration post-hoc processing as part of their design, the reliable post-hoc identification of different types of structures is a challenge for images in the generated ensemble, similar to real-world medical image ensembles. Minor adaptations to the SOM design might at least partially alleviate this issue and aid in designing robust and meaningful evaluation frameworks.

Implicit spatial context arises from stochastic interactions of individual structures; an example of implicit contextual assessment was demonstrated for the Voronoi-SCM in the previous chapter (see subsection 3.4.2). In case of the Voronoi SCM, although only a few parameters were described in the design of the SCM, other features that implicitly co-occur can be tested canonically. If the range of a certain feature is changed, a corresponding change in the values of co-occurring features might occur to maintain the identity of the realization. Thus, the features may not necessarily be independent of each other. This becomes more obvious in complex objects, and is the basis of designing evaluations for DGMs trained on stochastic models of *objects*, i.e, SOMs.

A purposefully designed dataset based on an SOM together with an evaluation framework based on the concept of implicit context are proposed in this chapter. This constitutes the second of the two evaluation frameworks proposed in this thesis. This framework was deployed towards a public Grand Challenge, which called for the submission of the "best-trained" DGMs, which were then ranked and individually analyzed for the presence of artifacts. Thus, the Grand Challenge enabled an evaluation of different DGM approaches based

on modern, state-of-the-art DGMs. A brief overview of the Grand Challenge is presented in the next section.

4.3 DGM-Image Statistics Challenge Overview

While several studies [34, 92, 149, 187, 188] have provided valuable insights into the performance of DGMs, the need for more widespread application-relevant assessments of DGMs in the field of medical imaging has been well-established.

To address this need and promote meaningful assessments and refinements of DGMs for medical imaging applications, we proposed a public Challenge: the Deep Generative Modeling for Learning Medical Image Statistics Challenge, or the DGM-Image Statistics Challenge for short. Each year the American Association of Physicists in Medicine (AAPM) issues a call for Grand Challenges in order "to assess or improve the use of medical imaging in both diagnostic and therapeutic applications". Our proposed Challenge was accepted and hosted by the AAPM under this call. The DGM-Image Statistics Challenge invited participants to develop or refine generative models that can accurately reproduce image statistics that are important and relevant to medical imaging applications, including the evaluation of imaging systems as well as for use in the training and testing of AI/ML algorithms. Through the DGM-Image Statistics Challenge, the following were made available: a dataset, a standardized evaluation procedure, and a benchmark for evaluating future generative models for medical image synthesis. A description of the DGM-Image Statistics Challenge framework and a reporting and discussion of the results are provided in this Chapter.

The DGM-Image Statistics challenge was unique in its focus on the development and evaluation of DGMs for creating ensembles of SOMs that could serve as inputs to a simulated medical imaging system, and could be evaluated via implicit contextual assessments. This Challenge probed the degree to which suitable SOMs might be synthesized via generative methods. In this direction, the DGM-Image Statistics Challenge aimed to facilitate the development of domain-appropriate DGMs, as well as promote the domain-relevant evaluation of DGMs. The specific goal of this Challenge was to identify the best DGM trained on the provided dataset that could accurately reproduce certain image statistics, as identified from the training dataset. In addition, the DGM still had to produce perceptually realistic images

and avoid overfitting/memorization of the training data. A summary measure was derived from these statistics to rank submissions and identify a winner and a runner-up.

4.4 Methods

4.4.1 Methods: Challenge Logistics

The challenge was conducted in two phases. In the first phase, the participants were expected to submit 10,000 images generated from their trained DGM, and a written summary of their approach. In the second phase, a packaged (dockerized) implementation of the DGM was expected. The rules of the challenge allowed only the use of the provided training dataset and a pre-trained network (trained by us and made available as a starting point for participants) for model development. A computational constraint on the generation process was also specified: the developed DGM was to be capable of generating 10,000 images in under 12 hours on a Nvidia V100 GPU with 16 GB RAM.

In the first phase, the FID score and a memorization measure based on image cross-correlation were employed to rule out submissions with obviously poor visual quality, or those consisting of memorized image ensembles. In the second phase, the DGM implementation provided by the participants were employed to generate 10,000 images. The generated image ensembles were evaluated as described later in subsection 4.4.3 to yield a summary measure that was employed for ranking the submissions. In addition, the code and images from submissions in the second phase were manually validated in the second phase. A subset of the features employed for evaluation were employed to provide a public measure that was available to the participants for the duration of the challenge so that they could fine-tune their models based on feedback from this public measure.

Note that the logistical organization of the Challenge was performed by collaborators on this project and is not a contribution of this thesis. This thesis focuses on the design of the dataset, the evaluation framework employed for the Challenge, and the additional analyses of all submissions.

4.4.2 Methods: Data Design

Although anatomical realism was important in the design of training data for the challenge, some practical aspects such as the typical computational requirements of training DGMs on large images, as well as robustness of post-hoc evaluation were also considered. A previously published stochastic object model of the human female breast: VICTRE [48], was adapted for this grand challenge. From each 3D volume generated at a voxel resolution of 0.1 mm^3 via the VICTRE tool, fifteen equidistant 2D slices were extracted from the central third of the volume. Note that some of these slices could appear similar and were not regarded as independent training samples. Eight different tissues or structures were described by VICTRE within this sub-volume. Only 4 of those tissue types: skin, fat, glandular tissue, and ligaments were retained in the adapted version for the challenge. This choice of tissues was based on (i) the presence of all 4 tissue types in each 2D slice, (ii) generally distinct tissue properties for the chosen imaging modality: x-rays at 30 keV, and (iii) the structural variety provided by the tissue types. Thus, this choice of tissues contributed to the robustness and utility of the evaluation method.

Structures belonging to the four omitted tissue types (artery, vein, duct, and terminal duct lobular units) were replaced by glandular tissue, which was the most similar tissue in terms of attenuation coefficients. Each slice was then downsampled to size 512×512. Even with this downsampling, the thin ligament structures in the original slices were retained. The data dimensions were chosen based on compute requirements of training modern DGMs on large images [52, 72, 117], and the time window of the challenge. The downsampling process involved the following: (i) the breast region in an image was identified with a bounding box, (ii) the four chosen tissue types within the bounding box were separated into four distinct binary arrays such that each pixel location was foreground only for one tissue, (iii) for each of these arrays, the coordinates of the foreground pixels were transformed to match a 512×512 array with a centered breast region, (iv) the ligament array was skeletonized [189], (v) all tissue arrays were thresholded, and (vi) the resulting four binary arrays—with exclusive tissue identity at each pixel location—were combined to yield a single image. Next, tissue specific-intensity distributions were pre-defined such that the relative tissue attenuation properties were largely maintained in the grayscale intensity range of 0 to 255. An exception was made for ligaments and skin tissues, which had very similar attenuation properties. The grayscale distributions for these two tissues were slightly adjusted to enhance their separability and aid post-hoc analyses. The final four tissue-specific intensity distributions were specified as

distinct Beta distributions:

$$\begin{aligned} t_{\text{fat}} &\sim 60\,X + 52, \text{ where } X \sim \text{Beta}(\alpha = 2, \beta = 4), \\ t_{\text{glandular}} &\sim 96\,X + 128, \text{ where } X \sim \text{Beta}(\alpha = 4, \beta = 2), \\ t_{\text{skin}} &\sim 16\,X + 228, \text{ where } X \sim \text{Beta}(\alpha = 3, \beta = 3), \\ t_{\text{ligaments}} &\sim 16\,X + 232, \text{ where } X \sim \text{Beta}(\alpha = 3, \beta = 3). \end{aligned} \tag{4.1}$$

The corresponding distributions for the four tissue types are shown in Figure 4.1.

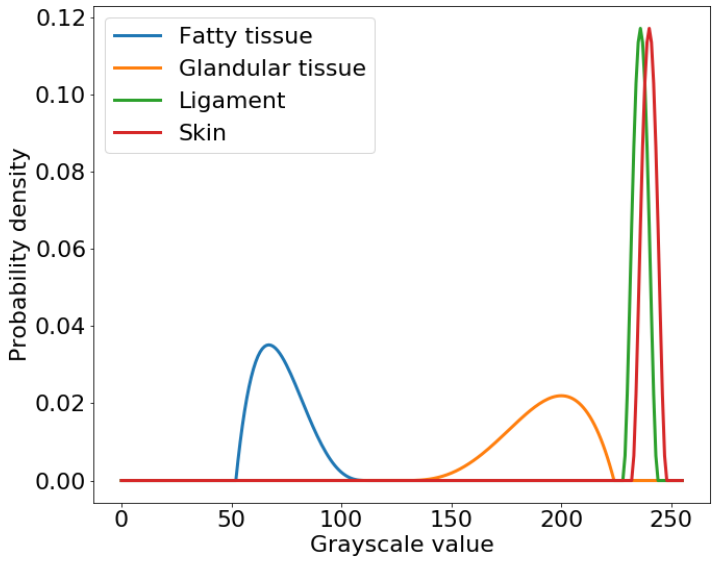


Figure 4.1: Tissue-specific intensity distributions in the adapted VICTRE breast phantom. Note that the fatty and glandular tissue distributions are distinct; this aids their segmentability. Although the intensity distributions of ligaments and skin have overlap, the expected locations of these two tissues are in the breast region are clearly different.

Next, variates from these intensity distributions were assigned to appropriate tissue locations as follows. Four arrays of size 512×512 were generated from each distribution for a single 2D slice. A texture was imposed on each array via a Gaussian filter with smoothing parameter $\sigma = 0.8$, similar to the process described in [190]. This resulted in slightly correlated pixels, thus, generating a prescribed texture, which could be then tested as part of the evaluation framework. Note that the extent of Gaussian smoothing was chosen subjectively, and not to a known correlation in tissue attenuation. After Gaussian smoothing, the histograms of the resulting arrays were transformed to ensure that the prescribed intensity distributions

were maintained. Each of the four arrays was then masked according to the known tissue locations and combined to yield the final 2D image. Last, the dataset was cleaned by eliminating images where the breast boundary was not entirely contained within the image.

The training ensemble consisted of 108,000, 8-bit images of size 512×512, saved via lossless compression, and was made available to the participants after registration for the challenge. This training ensemble comprised four breast types, as determined by the Breast Imaging Reporting and Data System (BI-RADS) classification system [191]. Note that the ratio of fatty to glandular tissue determines the breast type. In accordance with population prevalence [191], the four breast types, namely, fatty, scattered, heterogeneous and dense, were represented in the ratio of 1:4:4:1 within the training dataset. However, this information was not explicitly provided to the participants during the challenge. The challenge dataset is now publicly available along with the breast type label for each image [192].

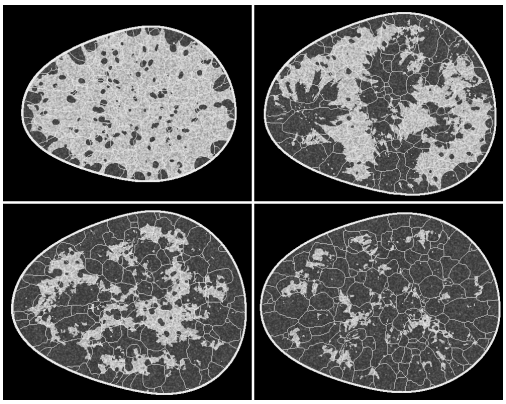


Figure 4.2: Sample images from the training dataset corresponding to four classes: dense (upper left), heterogeneous (upper right), scattered (lower left), and fatty (lower right). Class information was not provided explicitly to the participants.

4.4.3 Methods: Evaluation Strategy

From each DGM submission, an ensemble of 10,000 images was generated for evaluation. As described in subsection 4.4.1, the first stage of evaluation identified entries eligible for ranking via FID scores and a memorization measure. The memorization measure was the cross-correlation of binary masks representing the fatty-glandular tissue boundary, computed for all images in a DGM-generated ensemble against all images in the training ensemble. The boundaries between the two tissues were obtained via a series of simple filtering operations

followed by thresholding. An image in a generated ensemble was marked as memorized if the memorization measure exceeded a value of 0.9 on a [0, 1] scale. This threshold was determined as one standard deviation greater than the maximum value observed after calibrating this measure on a subset of 3000 randomly chosen images from the training ensemble against the remainder of the ensemble (about 108,000 images).

For the second stage of evaluation, all images in the training and DGM-generated ensembles were first segmented to obtain individual tissue regions. The segmentation process involved global thresholding that employed knowledge of the prescribed tissue-specific intensity distributions. Thus, a single image yielded the following tissue labels after segmentation: fatty tissue (F), glandular tissue (G), ligaments (L), skin (S). Features were then extracted from individual tissues (F,G,L,S) or the full breast slice (B), as indicated in the parentheses that follow the feature descriptor as follows: (i) texture features (B) [193, 194], (ii) morphological features (F, G, B) [189], (iii) skeleton statistics (L) [170], (iv) fractal features – fractal dimension and lacunarity (F, G, L) [195], (v) moments – raw, central, normalized, Hu, and their weighted versions (F, G, B) [189], and (vi) ratio of fatty to glandular tissue (B) [191]. For the computation of texture features, all data were binned to 64 gray levels, which were determined to be reasonable via the Freedman-Diaconis rule [196] for the training dataset. The feature sets above were chosen because they have been extensively employed for image classification and object recognition via conventional methods [170, 197–200]; we do not claim that these features are sufficient to describe diagnostic aspects of biomedical images.

Features that yielded multiple values for a single image, e.g., area of each disconnected fatty tissue region in an image, were summarized as: total count, mean, standard deviation, minimum value, maximum value and quartiles for each realization. In all, 3442 features were extracted from each slice over all feature families. Principal component analysis was performed on all features corresponding to the training data and each DGM-generated ensemble was projected into this principal component (PC) space after feature extraction. To obtain a baseline distribution corresponding to the training data, two data points represented as 10D vectors in the top-10 PC space were chosen at random and the cosine distance between them was computed; this process was repeated 10,000 times. For the DGM-generated ensemble, a similar computation was performed for one data point from the training ensemble and another from the DGM-generated ensemble, both represented as 10D vectors in the PC space of the training data. The two resulting distributions of cosine distances were then compared via the Kolmogorov-Smirnoff (KS) [201] test statistic. The procedure was repeated 1000

times on bootstrapped datasets to estimate the uncertainty in the KS statistic, and the resulting mean value of the KS statistic was employed to determine the final rankings in the challenge.

The evaluation pipeline described above was also employed for computing the public measure, and differed only in the choice of features. Only nine features derived from the intensity histogram and tissue areas were employed in the public measure computation.

For additional class-based analyses, not part of the ranking framework, a four-class classifier with a VGG-16 [202] backbone was trained on 5000 images per-class, for 400 epochs. Recall that the four breast classes correspond to dense, heterogeneous, scattered, and fatty breast types, and occur in the ratio 1:4:4:1 in the population. A validation set of 1500 images per-class was employed, and the model with the least validation loss was selected for inference. The training and validation datasets were distinct from the public dataset for the challenge. Calibration of the classifier on 3000 images per-class from the challenge dataset showed error-rates of 0%, 0.07%, 0.37%, 0.87% for the four classes. This classifier was employed on all images from the final submissions to predict a single class label for each image. The class prevalence in the generated image ensemble was then compared against the expected prevalence of the four breast classes from the training data, which was prescribed according to the population class prevalence.

4.4.4 Methods: Participants' Methods

The DGM approaches of submissions in the second phase are summarized in this section. To maintain participant anonymity, each group was assigned a random code consisting of letter and a number. Here onwards, the participants are referred to via this random code.

All participants employed/ adapted state-of-the-art DGM approaches based on GANs or diffusion models. Submissions based on GANs were: *K7*, *A8*, *V4*, *S4*, *J5*, and *H1*, whereas submissions based on diffusion models were: *D9*, *C2*, and *M3*, out of which the first two were based on conditional latent diffusion models [203, 204], and the last approach employed denoising diffusion GANs [205].

Extensive hyperparameter search with GANs, and without architectural modifications were undertaken by some groups such as *S4*, *J5* and *V4*. Results from these submissions demonstrated that hyperparameter tuning alone can provide substantial improvements in performance. However, these groups were not ranked in the top four submissions, indicating that this strategy alone may not be enough to achieve the best possible performance. The same groups along with one additional group (*H1*) also employed adaptive discriminator augmentation (ADA) strategies first proposed by Karras, *et al.* [206]. tThese included geometric transformations, intensity transformations, image flipping and rotation. These strategies led to some unexpected effects—artifacts arising from this strategy are described in the additional analyses in the Results section of this chapter.

Post-hoc processing methods were also employed by some groups. These included: (i) traditional image processing techniques, such as thresholding and filtering, employed by the groups *D9* and *H1*, and (ii) deep learning-based image superresolution methods, employed by the groups *K7*, *A8*, and *C2*. Thus, some fine-grained artifacts in the generated image ensembles could be eliminated before evaluation.

Interestingly, the top three submissions employed domain knowledge, i.e., information about the breast type, for conditional generation of images. The top-ranked submission conditioned on breast region area, as well as the fatty-to-glandular tissue ratio. The runners-up employed existing breast density classifiers trained on real-world data [207] and k-nearest neighbor (KNN) clustering to obtain labels for conditional generation.

4.5 Results

4.5.1 Participation Summary

The challenge received 58 submissions from 12 unique usernames. Out of the 12 unique submissions, six were from teams in the United States, two from India, and one each from Belgium, Brazil, Canada, and China. Split by sector, seven submissions came from academia or non-governmental organizations, two from the industry, and three from independent contributors or contributors with unknown affiliations.

4.5.2 Results: Overall Results

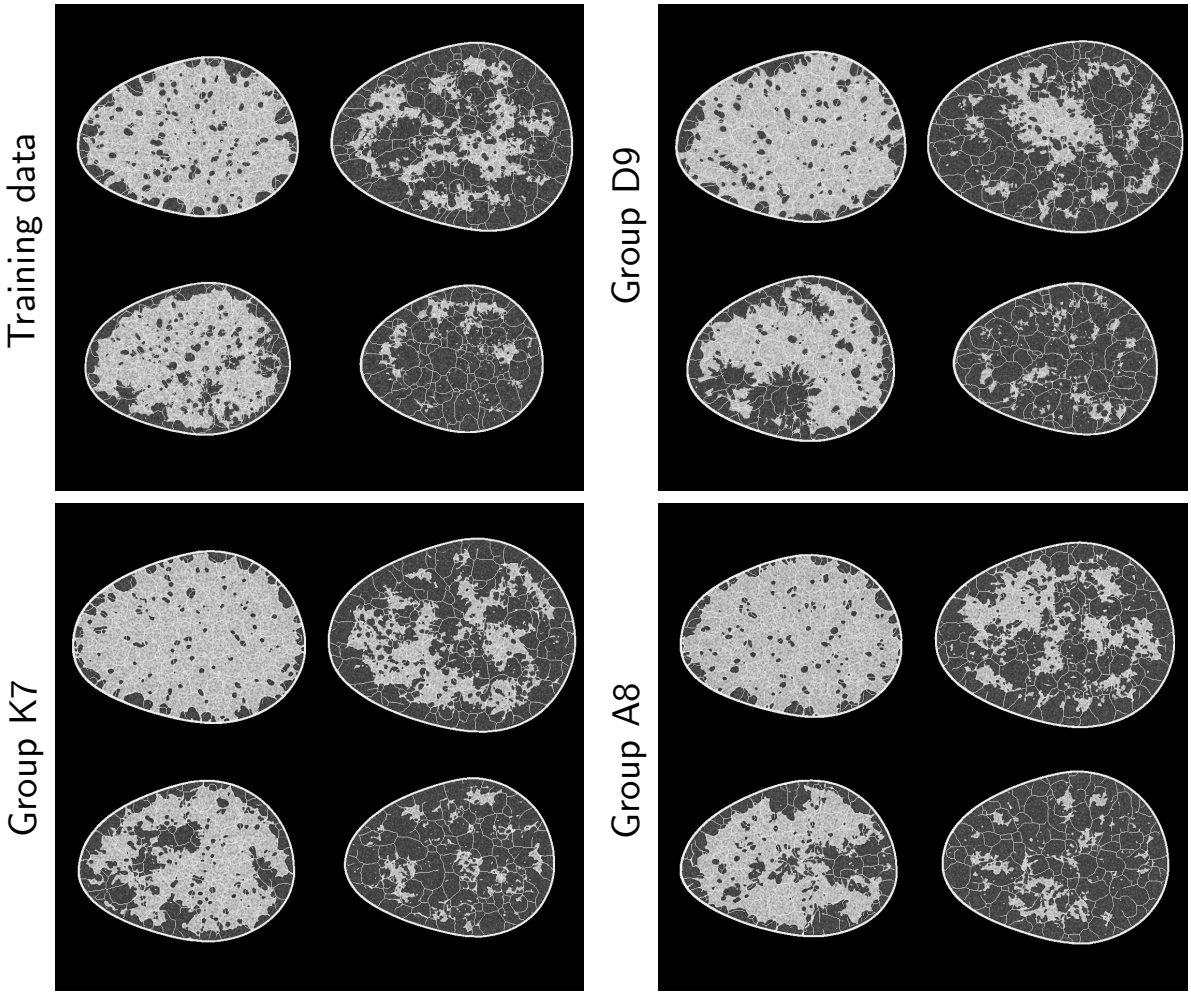


Figure 4.3: Images generated by the top three approaches alongside the images from the training data.

Images demonstrating high visual quality from the top-three approaches are shown in Figure 4.3. However, all three submissions also demonstrated artifacts as discussed in subsection 4.5.5.

The FID scores for the 12 unique submissions are reported in Figure 4.4 (left). Barplots of the memorization measure computed for 3,000 DGm-generated images from each submission are shown in Figure 4.4 (right). Submissions that produced images with memorization measure value > 0.9 were determined as being memorized. Images were manually checked to confirm

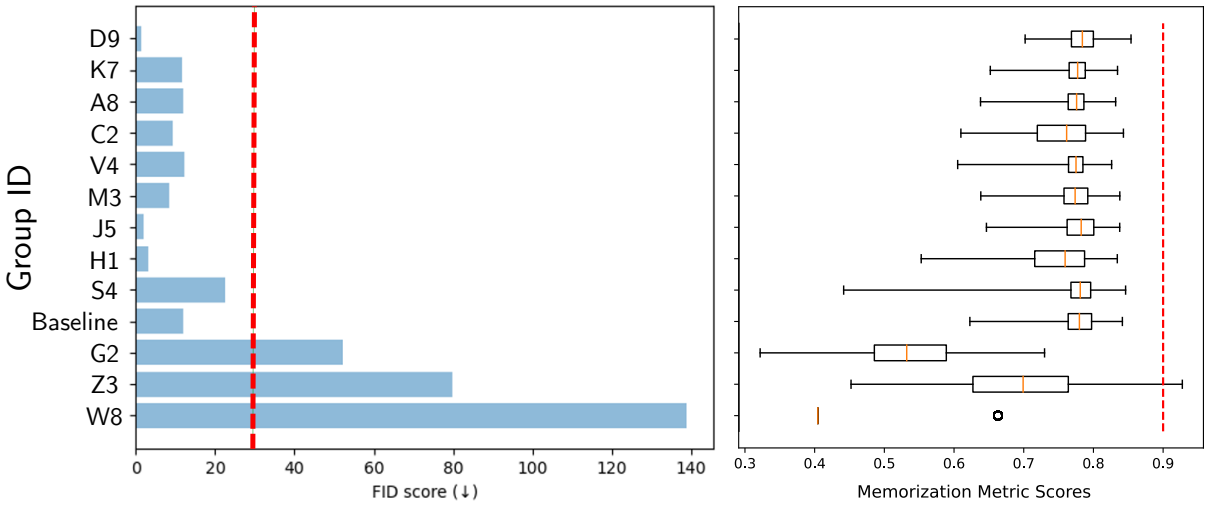


Figure 4.4: FID and memorization measure scores for the submissions alongside the FID and memorization measure scores of a baseline StyleGAN2 model trained in-house.

memorization. Note that only one submission was flagged for memorization. The FID score was also employed with a threshold of 30 to rule out images with obviously poor image quality. Three submissions were ruled out based on the FID and examples from these submissions are shown in Figure 4.5.

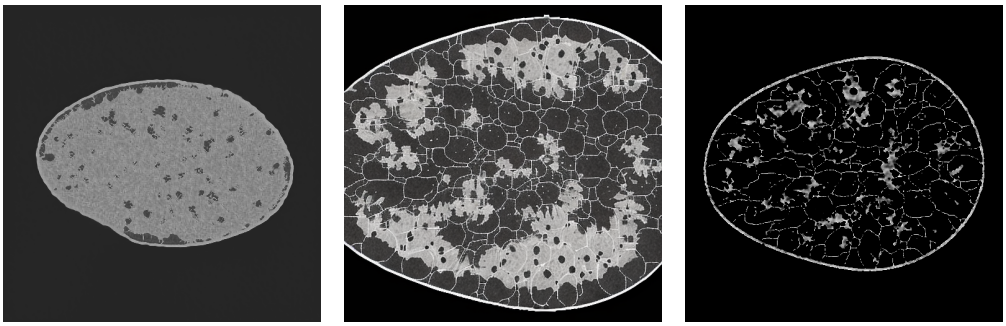


Figure 4.5: An image each from three submissions that were ruled out in the first stage of evaluation. The images in the left and center positions correspond to the submissions that did not pass the FID threshold, whereas the rightmost image corresponds to the submission that did not pass both the FID and the memorization thresholds.

After the first phase of evaluation based on the FID and the memorization measure, nine submissions qualified for the second stage. These were evaluated as described in subsection 4.4.3. Scatter plots of the first two principal components of all extracted features from

DGM-generated images are shown in Figure 4.6. Note that the distinctiveness of the training and generated ensembles increases with rank. Furthermore, the four clusters in the training data correspond to the four breast classes.

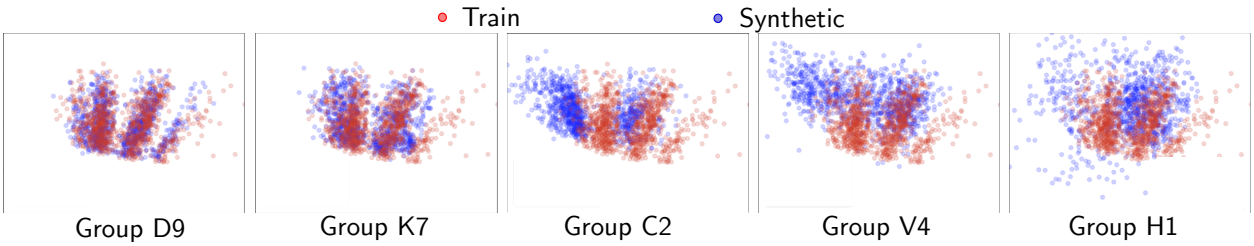


Figure 4.6: First two principal components of the features extracted from images from submissions ranked 1, 2, 4, 5 and 8.

Figure 4.7 (left) shows a bar plot of the final ranking measure for the nine submissions, and Figure 4.7 (right) shows the relationship of their FID-based rank to their rank based on the final ranking measure. It can be seen that for the submissions that were identified to be below a baseline FID-based threshold, the rankings based on the FID and the final ranking measure show poor correlation.

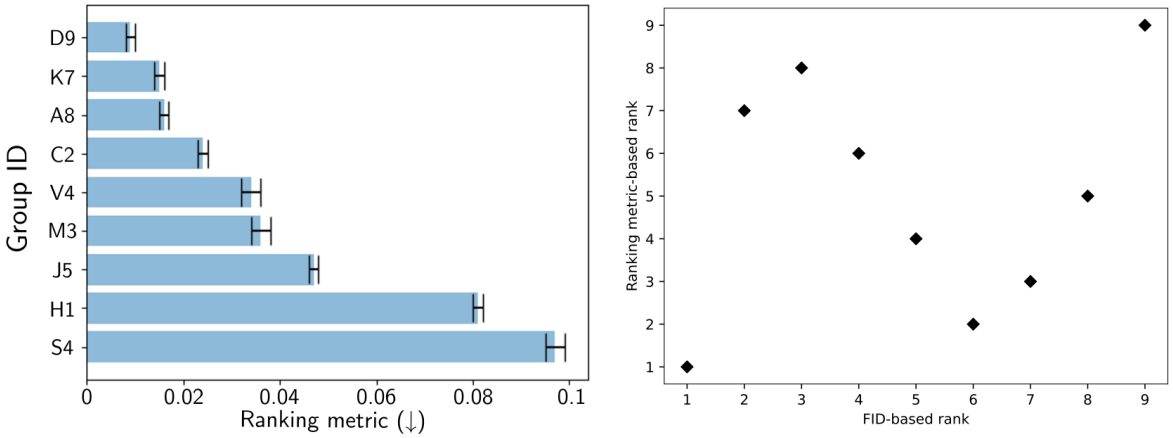


Figure 4.7: (Left) A barplot showing the ranking measure values for the 9 submissions that passed the FID-based threshold. Error bars indicate uncertainty. (Right) Rank of the submissions with respect to FID plotted against the rank of the submissions with respect to the ranking measure. Note that this plot only shows the submissions that have passed the FID-based threshold.

Results from the supplementary analyses of the final nine submissions are reported in the following three sub-sections. A representative subset of features from the original evaluation framework were employed for the additional analyses. Note that these analyses were not employed to determine the overall ranking and are presented only to provide additional insights into the different DGMs considered.

4.5.3 Results: Performance on Individual Feature Families

Table 4.1: Submission rankings based on individual feature families.
(F/G: fatty to glandular tissue ratio, Moment: normalized image moments)

User	Overall	Texture	F/G	Moment	Morphology	Fractal	Skeleton
D9	**1**	8	3	**1**	**1**	3	**1**
K7	2	**1**	8	4	5	**1**	3
A8	3	2	4	3	6	2	2
C2	4	4	**1**	2	8	8	4
V4	5	7	2	7	7	9	9
M3	6	9	9	6	3	7	6
J5	7	3	6	5	2	4	7
H1	8	5	5	8	4	5	8
S4	9	6	7	9	9	6	5

As described previously, to determine the overall ranking for the Challenge, all features were weighted equally. Rankings for *individual* feature-families are reported in Table 4.1. The overall top 3 submissions also ranked between 1 and 3 for several individual feature-families. The best submission performed remarkably well on most feature-families, except on the texture features. On the other hand, some lower ranked submissions were ranked high for a single feature-family (e.g., *J5* on morphological features, *C2* on F/G ratio). Thus, the choice of the "best submission" may vary based on the image statistics that are deemed important to a specified diagnostic task.

4.5.4 Results: Class-based Analyses

Class-based analyses were performed on the final submissions to gain insights into the composition of the generated image ensembles. The results are reported in Table 4.2. Recall that class information was not made public in the challenge.

Table 4.2: Class-based analyses of submissions. Expected class prevalence (%) from the training data is 10, 40, 40, 10 for the four breast types: fatty, scattered, heterogeneous, and dense, respectively. Class density and coverage is expected to be approximately 1 for all classes.

User	Rank	Class prevalence (%)	Class density	Class coverage
D9	**1**	10, 41, 39, 10	1.0, 1.0, 1.0, 1.0	0.9, 1.0, 1.0, 1.0
K7	**2**	10, 40, 40, 10	1.0, 1.0, 1.0, 1.0	0.0, 0.8, 0.9, 0.8
A8	**3**	10, 40, 40, 10	0.0, 1.0, 1.0, 1.0	0.0, 0.8, 0.9, 0.7
C2	**4**	3, 30, 55, 12	0.0, 0.7, 0.6, 0.2	0.0, 0.2, 0.4, 0.1
V4	**5**	9, 39, 35, 17	0.0, 0.3, 0.5, 0.2	0.0, 0.3, 0.3, 0.2
M3	**6**	12, 46, 38, 4	0.0, 0.2, 0.3, 0.3	0.0, 0.1, 0.2, 0.1
J5	**7**	11, 40, 40, 9	0.0, 0.7, 0.8, 0.7	0.0, 0.2, 0.7, 0.6
H1	**8**	11, 43, 35, 11	0.0, 0.4, 0.5, 0.2	0.0, 0.3, 0.4, 0.3
S4	**9**	14, 42, 40, 4	0.0, 0.7, 0.5, 0.5	0.0, 0.6, 0.6, 0.4

For all generated images, class was determined via the four-class classifier described in subsection 4.4.3 Most submissions demonstrate class prevalence similar to the training data: 10%, 40%, 40%, 10% for fatty, scattered, heterogeneous, and dense breast types respectively. Only two submissions (*S4*, *M3*) demonstrated instances of mode collapse with class prevalence below 4% for one class. However, class-wise density, and coverage [139] computed in the top-2 PC-space of features from all families, were rarely equal or perfect (approximately 1) across classes and submissions. Note that density and coverage are measures indicative of fidelity and intra-class diversity respectively [139]. Often, at least one class had nearly zero coverage, despite demonstrating ensemble prevalence similar to the training data. That is, although many realizations belonging to a class were generated, together, they did not capture the expected diversity within that class. Thus, although a generated ensemble may seem to replicate the class prevalence in the training data as determined by a forced-choice classifier, the fidelity of several class-specific features may still be suspect.

4.5.5 Results: Analysis of Artifacts

Visual inspection of all generated image ensembles revealed artifacts arising from ligament skeletons, morphology, and texture. Most importantly, some artifacts were not unique to a single submission, but were observed across multiple submissions, suggesting a correspondence to DGM approaches. All submissions demonstrated artifacts in ligament formation.

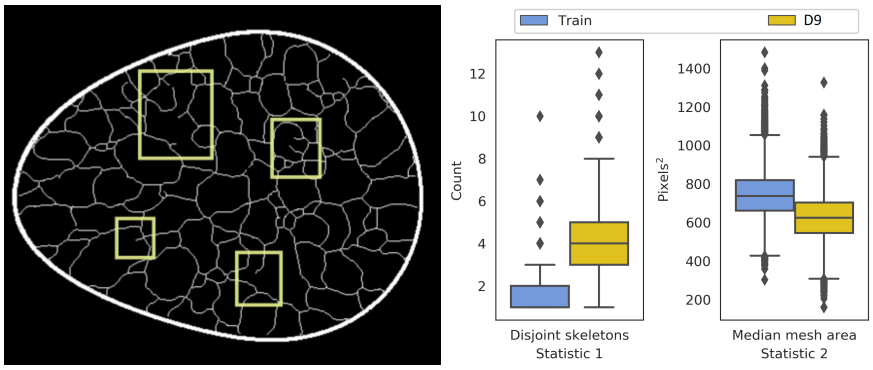


Figure 4.8: A thresholded sample image (left) from the top-ranked submission (*D9*) demonstrated clear breaks in ligament connections (yellow boxes). These artifacts were also reflected in two statistics: number of disconnected skeletons per-image, and the median area of bounded regions within an image. The boxplots correspond to 10,000 images each, from the training dataset, and the top-ranked submission.

Three kinds of artifacts were observed in the ligament skeletons: (i) sharp breaks in ligaments, (ii) gradual breaks in ligaments, blending into the background, and (iii) "ligament sticking", i.e., constant ligament structure across multiple images. The three artifacts are referred to as "break", "blend", and "stick" respectively in Table 4.3. The top-ranked submission demonstrated the first artifact, which was also captured in two statistics from the evaluation framework: (a) the number of disconnected skeletons per-image, and (b) the median region area over all regions in an image (see Figure 4.8). Thus, even the best submission was not perfect. The same artifact was also observed in another submission (rank 6), which also employed a conditional diffusion modeling approach like the top-ranked submission. The second artifact (see Figure 4.9) was observed across all other submissions, indicating that the generation of ligaments was not a trivial task for typical DGMs. Ligament sticking artifacts (see Figure 4.10) were observed in two of the final nine submissions. Note that this artifact was a genuine effect of the DGM, and not a user-defined feature; ligament sticking was also observed in some of our own experiments with DGMs.

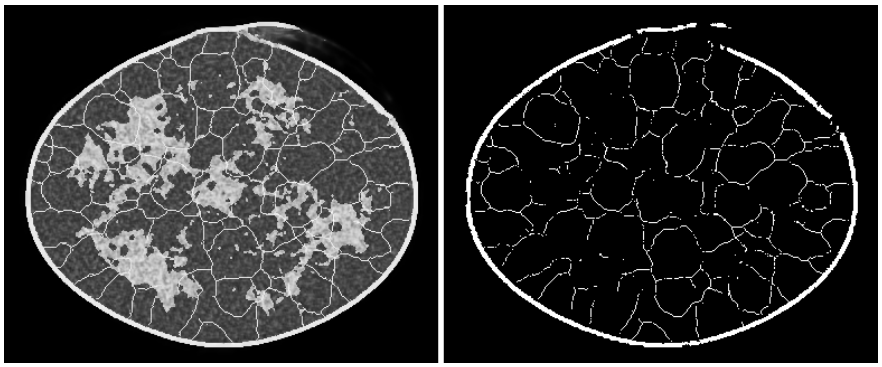


Figure 4.9: Two kinds of artifacts are demonstrated in a sample image from a submission: (i) broken boundary (left), and (ii) large, smooth breaks in the ligament skeleton (right).

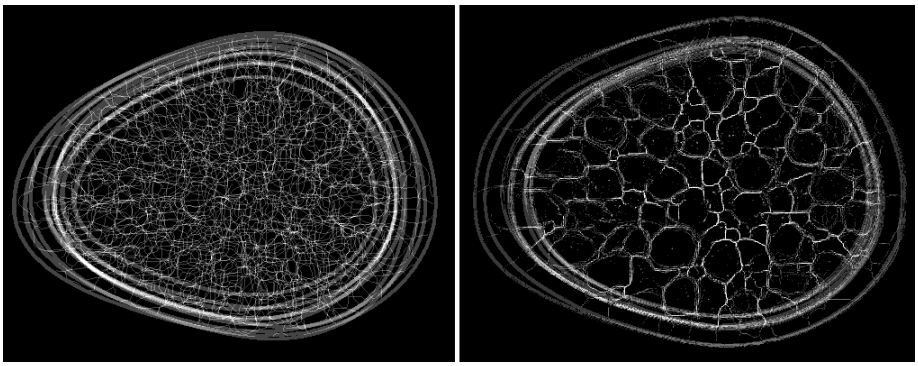


Figure 4.10: Some submissions demonstrated ligament sticking, that is, nearly constant ligament structure across images. Summation of 10 thresholded images of ligaments from the training data (left) demonstrates the expected randomness in ligament structure across realizations. This variation was clearly absent in similarly processed images (right) from a submission demonstrating this artifact.

Morphological artifacts of three types were observed across submissions: (i) broken boundaries, (ii) images flipped on the vertical axis, and (iii) malformations in the "burst"-like features in heterogeneous breast type. The three artifacts are referred to as "boundary", "flip", and "bursts" respectively in Table 4.3. Breaks in the boundary (see Figure 4.9), of varying degrees, were observed in three out of nine final submissions. To quantify the ensemble error rates of breaks in the boundary, the convexity [208], i.e., the ratio of the perimeter of the convex hull of the object to the perimeter of the object was computed, and a threshold of 0.9 was chosen after visual calibration. Within the three submissions that demonstrated breaks in the boundary, the ensemble error rate was observed to be 0.5 to 3% based on the convexity threshold (<0.9). All images from the training dataset were above this threshold. Four other submissions demonstrated instances of images ($< 0.1\%$ of the ensemble) where

the breast region was large enough that it was cut off by the image boundary; these submissions are not flagged for the "boundary" artifact in Table 4.3. The second morphological artifact: flipped images, comprised nearly *half* of the image ensemble for two submissions; this was likely an effect of enabling rotational augmentation during DGM training. The third morphological artifact ("bursts") was specific to the heterogeneous breast type. The characteristic burst-like patterns in the heterogeneous breast type were incorrectly formed in images from five submissions. Examples from submissions ranked 2, and 3, are shown in Figure 4.11; this artifact was captured in the lacunarity statistic computed on 3000 images classified as heterogeneus breast type, from each submission.

Artifacts in texture were observed in the foreground as well as the background regions. These artifacts are referred to as "fore", and "back" respectively in Table 4.3. Obvious foreground artifacts were clearly visible in all except the top three submissions. Two such examples are shown in Figure 4.12. Similarly, background artifacts (see Figure 4.13), i.e., non-zero background pixels, were found in all except two submissions; however, these were not visually obvious. The mean fraction of per-image, non-zero background ranged from 11% to 62% over the ensemble for the bottom three submissions for this artifact. However, the ensemble mean of the per-image mean grayscale value over the non-zero background pixels was below 6 for all submissions, and the ensemble standard deviation was at most 18. Recall that the original training data was 8-bit, i.e., demonstrated pixel values from 0 to 255. Thus, the background artifacts were well below the least values in the foreground, and hence, could be eliminated via thresholding. The two submissions that did not have any background artifacts reported employing post-processing techniques on generated images.

A summary of all artifacts across the final nine submissions is given in Table 4.3. A check mark indicates that at least one instance of the artifact was observed in the ensemble. Note that this list of artifacts is not comprehensive, but is indicative of some common artifacts across submissions. Other less visually obvious artifacts may have been captured in the evaluation framework but not described in this section.

Last, we assessed the diversity of an ensemble. From our knowledge of the VICTRE phantom, we intuitively expect ergodicity, which, here, means that pixel locations should not be strongly tied to tissue identity. Therefore, we computed the semivariance [209] of the mean image, which was obtained by taking the per-pixel mean over the ensemble (Figure 4.14). In all but two submissions that were ranked 1 (see Figure 4.14) and 6 (not shown), certain

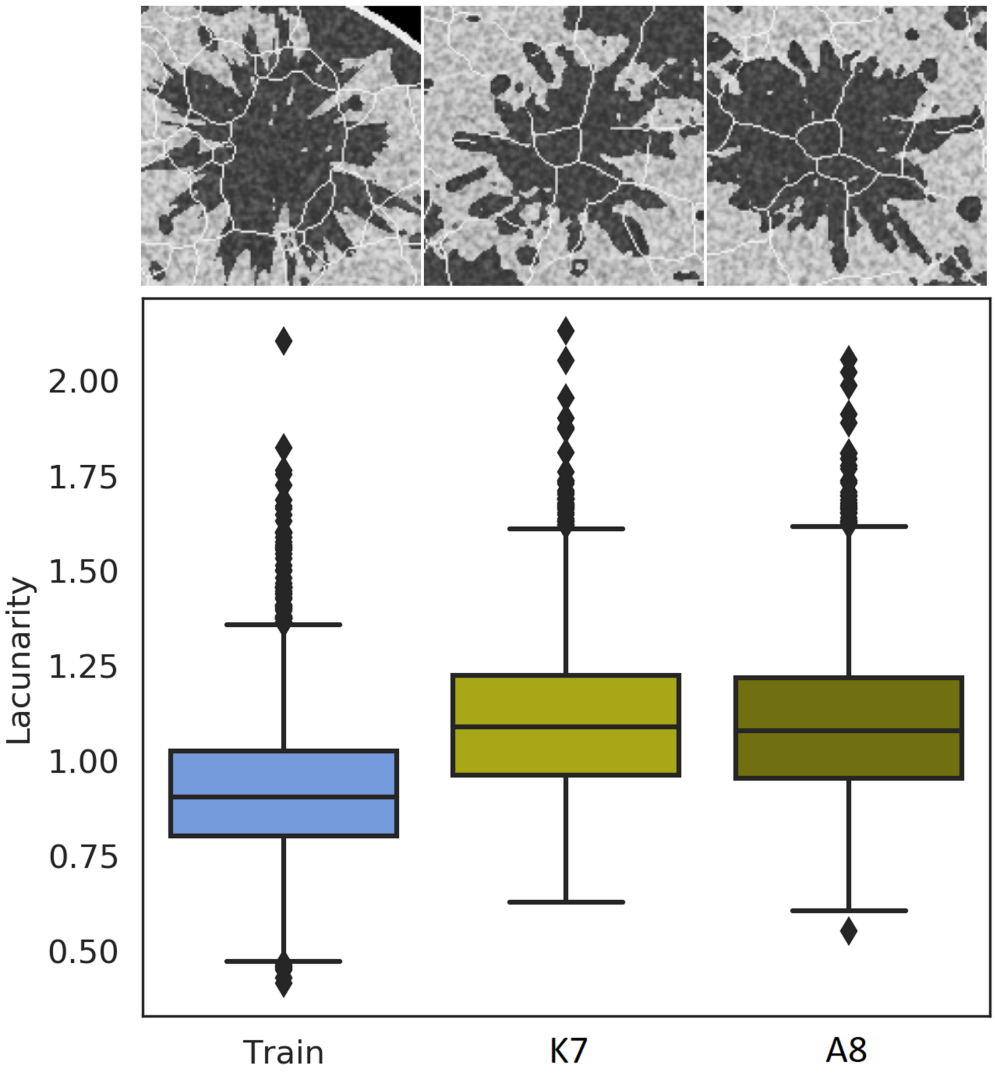


Figure 4.11: The characteristic "burst" pattern, typically observed in the heterogeneous breast type, was malformed in several final submissions. Top: An example from the training data (left) shows sharp, splinter-like features, while examples from the second-, and third-ranked submissions (center and right respectively) demonstrate rounded, splatter-like patterns. Bottom: The differences are captured via the lacunarity statistic.

pixels were preferentially allocated certain breast tissue types. Examples of submissions where the expected per-pixel means were not matched are also shown in the corresponding mean images computed over 10,000 images from the respective ensembles (Figure 4.14). This indicates lower diversity than the training ensemble. Note that the two submissions with ensemble diversity comparable to the training dataset were both conditional latent diffusion models. Because the breast region in the mean images is approximately the same size and

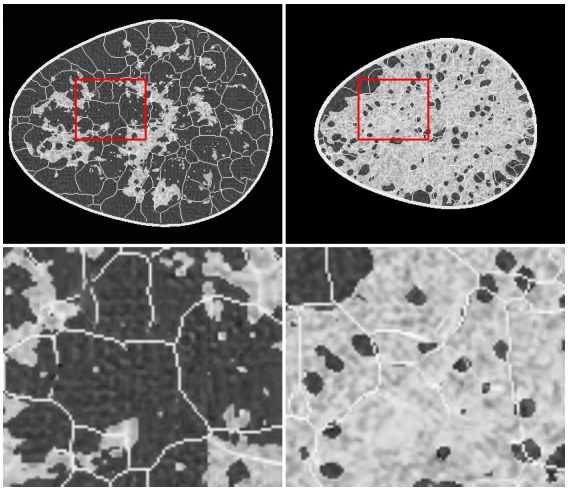


Figure 4.12: Sample images from two different submissions demonstrating texture artifacts in the foreground region. Gridding, or checkerboard artifacts (left), and "eddies" (right) were observed.

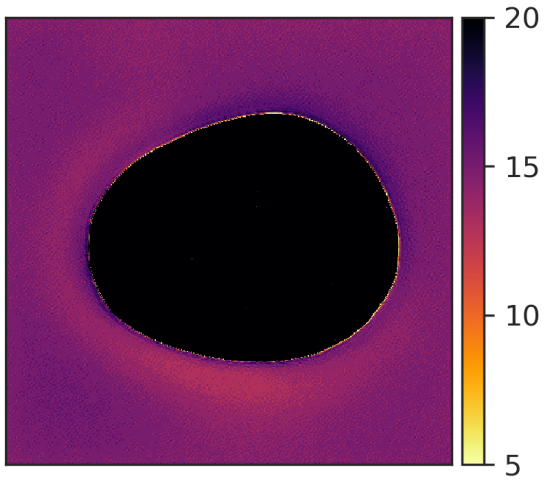


Figure 4.13: Sample DGM-generated image demonstrating artifacts in the background, which should ideally be constant at zero. Contrast is adjusted for display.

Table 4.3: Overview of artifact types visible in final submissions.

User	Rank	Skeleton			Morphology			Texture	
		break	blend	stick	boundary	flip	bursts	fore	back
D9	1	✓							
K7	2		✓				✓		✓
A8	3		✓				✓		✓
C2	4		✓		✓	✓	✓	✓	✓
V4	5		✓	✓		✓		✓	✓
M3	6	✓					✓		✓
J5	7		✓					✓	✓
H1	8		✓		✓	✓	✓	✓	
S4	9		✓	✓	✓		✓	✓	✓

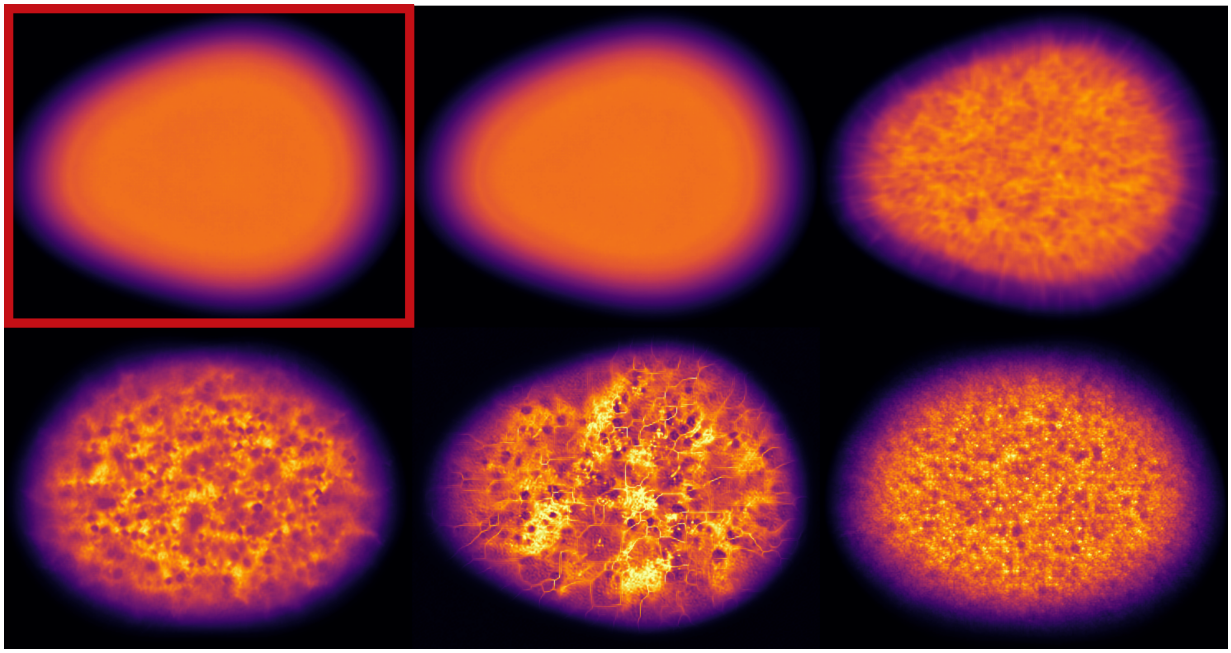


Figure 4.14: Ensemble mean images over 10000 randomly selected slices from the training data (highlighted in red), and five submissions ranked 1, 2 (row 1, center, and right), 4, 5, and 8 (row 2, left to right), demonstrate positional preference in tissue locations. Ideally, the mean image should appear similar to the mean image of the training data, i.e., nearly constant in the central region and smoothly decaying along the average boundary of the breast region. Distinct structure within the breast region is clearly visible in all except the first-ranked submission, and is indicative of pixel-specific bias in tissue generation.

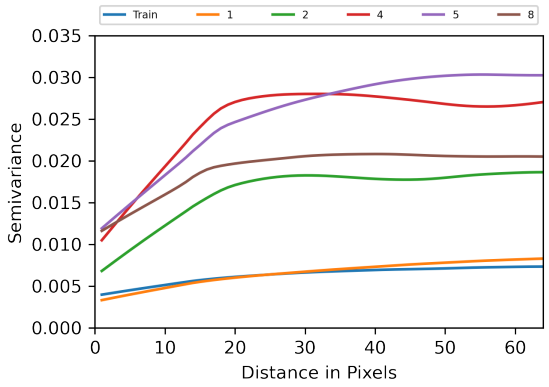


Figure 4.15: Semivariance of the mean images in Figure 4.14. The "sill" of the semivariance, i.e., the value at which the curve flattens, determines ensemble diversity. Low values of the sill indicate high ensemble diversity, and vice versa. Legend indicates overall rank of the submission.

centered, the visible differences can be summarized via their semivariances [209]. The semivariance is a spatial statistic that plots the differences in the intensity values of a random pixel-pair (Y-axis), binned by the distance between the two pixels (X-axis), over a large sample of randomly chosen pixel-pairs. Pixel-pairs that are closer together are represented closer to the origin on the X-axis and pixel-pairs with large separation are farther away from the origin on the X-axis. Thus, when mean images are nearly constant over a large region of the image, the semivariance would appear as a flat line with its slope close to zero. In other words, the long term constancy of the semivariance is indicative of high ensemble diversity. As seen in Figure 4.15, the top-ranked submission is almost as diverse as the training dataset. Furthermore, when the mean image is not constant and exhibits clusters of pixels of certain characteristic size, this effect is also captured in the semivariance at the corresponding distance. The "sill" of the semivariance, the value at which the curve flattens, is indicative of the level of ensemble diversity in this analysis. If the curve flattens at a low value of distance, it implies high ensemble diversity, and vice versa. The submissions ranked 2, 4, 5, and 8, clearly demonstrated lower ensemble diversity than the training dataset, and also differed among themselves in terms of the length-scales of pixel correlations.

4.6 Discussion and Conclusion for Chapter 4

Results from the Challenge highlight the fact that a single number cannot represent all the different ways in which DGMs can make errors. Multiple kinds of evaluations/figures-of-merit are required to assess different aspects of image quality of DGM-generated image ensembles.

In the designed evaluation framework, various feature families were employed towards this goal. These feature families are well-established in the image processing literature and were extensively employed before deep learning methods became popular for image analysis and classification. Most importantly, the employed features are interpretable, i.e., the formula of a feature can be tied to visually apparent features in the generated images. This also enables the detection of artifacts in an objective manner.

The design of objective evaluation frameworks for biomedical images such as the VICTRE SOM is challenged because clinical knowledge, e.g., anatomy, physiology, tasks, may not be directly and uniquely mapped to numerical observers. Given this challenge, one approach to objective evaluation is via image statistics. Different sets of statistics might be relevant to different tasks. Hence, a broad range of feature families was chosen for evaluation. However, it is important to note that: 1) different DGMs may demonstrate superior performance for different feature families, as observed in Table 4.1 and 2) the set of features employed in the proposed framework may not generalize to a dataset that does not look like the VICTRE phantom. For example, skeleton statistics were chosen as a feature family because ligaments are present in images, but the same statistics may prove to be a confounding factor if no ligament-like structure was present in the images.

Another important aspect to note is that the data employed in this Challenge was not corrupted in any way by the action of an imaging system. Any non-ideal imaging system might worsen image quality, or at least impact the range of the studied statistics. Given the current framework for the Challenge, all artifacts observed in the generated images were unequivocally caused by the DGM. The proposed framework allowed us to study the effects of DGMs alone, without being conflated with the effects of an imaging system. In future, if a DGM achieves excellent image quality on this "clean" dataset, it then becomes a candidate to be tested under conditions describing a more realistic imaging system.

The choice of the VICTRE phantom to create the dataset for this Challenge had several advantages. First, this phantom has been employed for virtual imaging trials, and hence, its diagnostic relevance is well established. Second, a high variety of structures is present in this phantom; this enables the assessment of various DGM capacities. For example, the phantom contains: (i) irregular shapes and edges resulting from the packing of the fatty and glandular tissues, (ii) regular shapes such as the ellipsoidal breast region boundary, and (iii) thin, long structures that constitute the ligaments. A DGM is expected to learn to create all different kinds of structures and place them correctly. Third, the various features in the VICTRE phantom occur at multiple scales. That is, some structures such as ligaments may have a width of a few pixels, whereas the breast region itself spans several hundred pixels. Thus, all the three aspects of the VICTRE phantom provide enable a wide range of tests based on domain knowledge.

In the design of the VICTRE phantom, the grayscale intensity values are generally consistent with the relative attenuation coefficients of all tissues. Furthermore, unlike real data, each tissue was prescribed an intensity distribution which ensured tissue segmentability and provided an additional feature for assessing DGMs.

There are some limitations of the proposed evaluation framework. Because image statistics are computed on segmented tissues from each generated image, accurate segmentation of tissues is important. If, for example, the intensity distributions of all tissues were shifted as compared to the prescribed distribution, all tissues may not be perfectly segmented, and hence, the image statistics would include the effects of the incorrect tissue segmentation. If the proposed evaluation framework were to be employed to study a DGM, the network could be afforded some intensity transformation consistent with visually distinguishable tissues before image statistics are extracted. Another limitation of this framework occurs in class-wise analysis. A DNN classifier trained on the original dataset is employed to predict class from generated images. Even if a DGM generates images with extremely poor image quality, the classifier allocates a class to these images. One way to alleviate this issue is to remove outliers before class prediction. Alternatively, other measures that assess class distributions in an ensemble could be employed together with the classifier to provide a more comprehensive evaluation of class fidelity and diversity. Last, the range of some image statistics can be tied to class identity. Hence, a class-conditioned analysis of various feature families might provide more insight into DGM capacities. A major finding from this Challenge was that different DGMs produce the same kind of image artifacts as summarized in Table 4.3. Some artifacts

are explainable or consistent with findings from other works, e.g., the image flipping artifact is a result of data augmentation methods during training, and the checkerboard artifact has been reported by several works as a potential result of certain upsampling operations in a DGM. Other artifacts such as the "bursts" could be characteristic of DGM architectures. Identifying artifacts characteristic to DGMs can enable the design of novel architectural solutions as well as post-hoc processing methods to eliminate imperfect images from generated ensembles.

In conclusion, an evaluation framework based on (i) a complex SOM describing anatomy, and (ii) the implicitly arising contextual features, was developed in this chapter for assessing DGMs for contextual hallucinations.

Thus, as described in chapter 3 and chapter 4, two model-agnostic evaluation frameworks for the assessment of reproducible spatial context were designed. In the next chapter, both frameworks are deployed to gain insights into a state-of-the-art DGM approach that has been reported to produce images of excellent visual quality.

Chapter 5

Assessment of a Diffusion Generative Model for Reproducible Context

"It is easy to destroy, but it is hard to create." - Pearl S. Buck

5.1 Overview

Diffusion models have recently emerged as a popular family of deep generative models (DGMs), particularly due to the excellent visual quality of generated images and a strong theoretical foundation. The top-ranked model in the Grand Challenge described in the previous chapter was also a kind of a diffusion model. Furthermore, diffusion models in the Grand Challenge demonstrated clearly higher ensemble diversity than the other submissions. This suggests that the diffusion model paradigm may possess a greater capacity to reproduce domain-relevant information as compared to other state-of-the-art approaches, however, this notion remains unexplored.

The fundamental idea behind diffusion models is that if a diffusion process can map the training data to standard Gaussian noise (over several discrete transitions), and the reverse mapping can be learned, then, new samples similar to the training data can be generated from Gaussian noise. This idea is inherently different from the previously popular approach: generative adversarial networks (GANs), which employed adversarial training to implicitly learn the distribution of the data by judging the quality of generated random variates, but do not seek to estimate the distribution itself. In the literature, it has been claimed that one class of diffusion models—denoising diffusion probabilistic models (DDPMs)—demonstrate superior image synthesis performance as compared to GANs. To date, these claims have

been evaluated using either ensemble-based methods designed for natural images, or conventional measures of image quality such as structural similarity. Because the perceptual image quality of images generated by DDPMs is extremely high, it is possible that domain-specific hallucinations often go unnoticed.

In this chapter, the evaluation frameworks proposed in the previous two chapters, i.e., tests of explicit spatial context (SCM-based evaluation) and implicit spatial context (SOM-based evaluation), are employed to investigate the ability of the DDPMs to reliably reproduce spatial context. Note that the DDPM employed in this work has more learnable parameters as well as computational budget than the top-ranked model in the previous chapter, and hence, is expected to provide a higher benchmark. The studies undertaken in this chapter reveal new and important insights regarding the capacity of DDPMs to learn spatial context. Notably, the results demonstrate that DDPMs hold significant capacity for generating contextually correct images that are 'interpolated' between training samples, which may benefit data-augmentation tasks in ways that GANs cannot. It was also observed that no generated ensemble from any diffusion model perfectly reproduced the expected spatial context.

5.2 Introduction

Significant advancements in DGMs have been achieved in the last few years [62, 80, 82, 117, 210]. Recently, a novel paradigm based on diffusion generative modeling [211] has been actively developed and explored in medical image research [4, 212–214]. The rapid and widespread adaptation of diffusion models for medical imaging applications has occurred due to the extremely high visual quality of images as reported in the computer vision literature [52,66], as well as the strong theoretical foundation of diffusion models [66], as compared to the previously popular DGMs: generative adversarial networks (GANs). This popularity of diffusion models is despite their higher sampling time than GANs [50]. The high visual quality and ensemble diversity attributed to diffusion models was also evident in our results from the Grand Challenge reported in chapter 4. Given the high visual quality of the images generated from diffusion models, domain-relevant errors may not always be captured via visual evaluations by non-domain-experts. Thus, the evaluation of domain-relevant contextual errors becomes even more important when domain-agnostic benchmarks of visual quality are satisfied.

Several works have claimed that diffusion models demonstrate superior performance in medical image synthesis as compared to GANs [213–217]. Muller et al. [34] explored the performance of diffusion models with respect to GANs for three different medical imaging modalities: eye fundus images, histology images, and chest radiographs. For the three modalities, they found that diffusion models generally demonstrated superior image quality and greater ensemble diversity. These claims are based on the FID score [45], precision and recall measures [49], and a classification task. They also report instances of contextual errors in GAN-generated images all three imaging modalities (refer Figure 1.3) e.g, incorrect number of optical disks in the eye fundus, checkerboard texture in the histology images, and incorrectly placed medical devices in chest radiographs. Occasional contextual errors in one imaging modality: chest radiographs, were also reported for the studied diffusion model. Other works based on diffusion models also employ versions of the FID score, precision and recall metrics, or SSIM to demonstrate the superiority of diffusion models over GANs in applications such as generation of histopathology images [218], and brain MRI [213]. The limitations of these measures have been discussed in a previous chapter (chapter 2). Recall that the relevance of some of these evaluation measures for medical image assessment has received limited attention [219], and thus remains largely unknown. Although contextual errors have previously been reported in GAN-generated ensembles [137, 220, 221], their occurrence in diffusion-model-generated ensembles has not been studied systematically. Thus, although innovations in DGMs, such as diffusion generative models, have been translated rapidly from computer vision to medical imaging research, their capacity to reproduce domain-relevant context is not yet established. In this chapter, the suitability of a diffusion model for medical imaging applications is tested via its capacity to reproduce prescribed spatial context via the two frameworks proposed in chapter 3 and chapter 4.

Three related formulations appear in the diffusion modeling literature: denoising diffusion probabilistic models (DDPMs) [51, 52, 81], score-based generative models (SGMs) [80], and stochastic differential equations (SDEs) [82]. The first formulation, DDPM [51, 52, 81], is chosen as the focus of this chapter. DDPMs are designed to estimate the probability density function of the target data distribution by learning to translate an initial noise distribution to the target data distribution over multiple intermediate steps that are modeled as probabilistic transitions. DDPMs also have high training stability and mode coverage. They are a popular choice for medical imaging applications ranging from medical image synthesis [4, 212–215] to image reconstruction [210, 222–224].

In this chapter, a DDPM was first tested for its capacity to reproduce *explicit* context via the previously established test bed of stochastic context models (SCMs) (refer chapter 3). Recall that the SCMs represent attributes relevant to medical imaging, in a readily interpretable manner and without anatomical constraints. Next, the DDPM was tested for its capacity to reproduce *implicit* context via the adapted version of a previously published stochastic object model (SOM) that describes anatomical constraints [48] (refer chapter 4). Per-image, contextual errors in DDPM-generated ensembles were then quantified to provide a measure of the potential suitability of DDPM to medical imaging tasks involving similar contextual attributes. This evaluation approach was also employed to gain insights into other adaptations of the DDPM formulation as well.

5.3 Background: Denoising Diffusion Probabilistic Models (DDPM)

In the DDPM framework [52], a small quantity of Gaussian noise is gradually injected into an input image (sampled from a real data distribution) $\boldsymbol{x}_0 \sim q(\boldsymbol{X}_0)$ over t time steps to eventually obtain the degraded image $\boldsymbol{x}_t$. Recall that the dimensionality of $\boldsymbol{x}_0$ and $\boldsymbol{x}_t$ is the same as that of the flattened image (refer subsection 2.2.2). Over a sufficiently large number of time steps T, a sample $\boldsymbol{x}_T$ from a Gaussian distribution can be produced. The forward diffusion process is formulated as a Markov chain where $\boldsymbol{x}_t$ and $\boldsymbol{x}_{t-1}$ are related by the transition rule defined as:

$$\boldsymbol{x}_t = \sqrt{1-\beta_t}\boldsymbol{x}_{t-1} + \sqrt{\beta_t}\boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}) \tag{5.1}$$

$$q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) = \mathcal{N}\left(\boldsymbol{x}_t; \sqrt{1-\beta_t}\boldsymbol{x}_{t-1}, \beta_t \boldsymbol{I}\right). \tag{5.2}$$

Here, $\mathcal{N}(\mathbf{0}, \boldsymbol{I})$ is a multivariate Gaussian distribution with zero mean and identity covariance $\mathbf{I}$, $\beta_t \in (0, 1)$ is a parameter controlling the addition of noise, and $\boldsymbol{\epsilon}$ is a variable from a standard, multivariate Gaussian, representing noise.

The reverse diffusion process that maps $\boldsymbol{x}_T$ to $\boldsymbol{x}_0$ is also formulated as a Markov chain, where each step represents an incremental denoising of the data. The reverse transition probability

between $\boldsymbol{x}_{t-1}$ and $\boldsymbol{x}_t$ can be represented by a Gaussian distribution for a large T and small β_t:

$$q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t) := \mathcal{N}(\boldsymbol{x}_{t-1}; \boldsymbol{\mu}(\boldsymbol{x}_t, t), \boldsymbol{\Sigma}(\boldsymbol{x}_t, t)). \tag{5.3}$$

A neural network, which also takes the time-step as an input, is employed to approximate the reverse mapping by predicting the mean ($\boldsymbol{\mu}$) and covariance ($\boldsymbol{\Sigma}$) for all reverse diffusion steps.

Note that the diffusion model described here is a latent-based model, wherein $\boldsymbol{x}_1, ..., \boldsymbol{x}_T$ are the latents. Following the rationale of training latent-based models as described in an earlier section (2.2.2), the distribution of the data may be approximated by its lower bound, i.e., the evidence lower bound (ELBO), and employed to train a latent-based generative model via variational inference.

Similarly, in case of DDPM, the variational lower bound (L_{vb}) was employed in the loss function to minimize the negative log-likelihood:

$$L_{vb} = \mathbb{E}_{q(\boldsymbol{x}_{0:T})}\left[\log \frac{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}{p_\theta(\boldsymbol{x}_{0:T})}\right] \geq -\mathbb{E}_{q(\boldsymbol{x}_0)}\left[\log p_{\boldsymbol{\theta}}(\boldsymbol{x}_0)\right], \tag{5.4}$$

where $p_{\boldsymbol{\theta}}$ is the network parameterization for the approximated reverse diffusion process, $\boldsymbol{\theta}$ represents the network parameters, and $\mathbb{E}_q$ represents expectation over q. The collection of data samples between time steps 0 and T is represented by $\boldsymbol{x}_{0:T}$, while the image samples between time steps 1 and T conditioned on the sample at time step 0 are represented by $\boldsymbol{x}_{1:T}|\boldsymbol{x}_0$. Here, the term in the center represents ELBO, which is greater than or equal to the negative log likelihood of the distribution of the training data. Recall that the ELBO serves as a bound for the likelihood of the data, which is often intractable in higher dimensions.

The bound can be reformulated as described in Dhariwal and Nichol [52]:

$$\begin{aligned} L_{vb} &= \log p_{\boldsymbol{\theta}}(\boldsymbol{x}_0|\boldsymbol{x}_1) \\ &\quad - \sum_{t=1}^{T} \mathrm{KL}(q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0) \,||\, p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)), \end{aligned} \tag{5.5}$$

where KL denotes Kullback-Leibler divergence. Here, the first term represents the likelihood at the first time step whereas the second term is equivalent to the bounds on the likelihood of all remaining time-steps, rewritten in terms of KL divergences (refer section 2.2.2). Note that $\mathrm{KL}(q(\boldsymbol{x}_T|\boldsymbol{x}_0)\,||\,p(\boldsymbol{x}_T))$ is omitted because it does not depend on $\boldsymbol{\theta}$.

The second term can be reparameterized further, as described by Dhariwal and Nichol [52] such that a DNN can be employed to predict the noise at each time-step. Once a network is trained to predict noise at each time-step, it can then be employed to represent the reverse diffusion process. In other words, the trained DNN can sample from the data distribution, given a random sample from a multivariate Gaussian distribution. Note that the DDPM approach is clearly different than GANs mainly in the following ways: (i) the dimensionality of the latent space is the same as that of the image for DDPMs and typically lower than the image for GANs, (ii) in the training process, the pixel co-ordinates are maintained for DDPMs but not GANs, which involve a decoding process from a lower dimensional latent vector, (iii) the data distribution is approximated in DDPMs but not in GANs. All three aspects potentially contribute to improved learning of long-range correlations. In addition, improved mode coverage and training stability has also been reported for DDPMs as compared to GANs [50, 52, 66].

5.4 Methods

5.4.1 Methods: Evaluation Frameworks

The evaluation frameworks presented in chapter 3 and chapter 4 were employed for the assessment of the DDPM. First, the three SCMs proposed in chapter 3 were employed for the assessment of prescribed spatial context. Then, the adapted breast phantom described in chapter 4, here onwards referred to as the VT-SOM, was employed for the assessment of implicit spatial context. Together, the four stochastic models encode a variety of contextual constraints. An overview of these constraints is provided in Table 5.1. All other aspects of the training data and the evaluation procedures were as exactly as described in the previous chapters.

Table 5.1: Overview of all stochastic context and object models in terms of the *per-image* contextual constraints explicitly prescribed in the model.

Constraints	A-SCM	V-SCM	F-SCM	VT-SOM
Prevalence	✓	✓	✓	✓
Intensity	×	✓	✓	✓
Texture	×	✓	✓	✓
Position	✓	×	✓	✓
Anatomy	×	×	×	✓
Multi-class	×	✓	✓	✓

Some sample images from all four stochastic models are shown in Figure 5.1 (SCMs) and Figure 5.2 (VT-SOM) for reference.

5.4.2 Network Trainings

Two popular diffusion models: Denoising Diffusion Probabilistic Model (DDPM) [52] and MedFusion—a kind of a latent diffusion model (LDM) [215] were employed in this work. The LDM approach partially alleviates the high compute requirements and sampling time of the DDPM by training a DDPM in the latent representation learned by a variational autoencoder. In other words, this approach reduces the computational burden by decreasing the dimensionality of the variables expected to be learned by a DDPM.

The DDPM was trained on all three SCMs and one SOM, whereas the LDM was trained on one SCM: V-SCM, and the VT-SOM. The DDPM has greater model capacity than the LDM and retains the original dimensionality of the data. Thus, the DDPM could potentially provide a kind of an upper bound in performance between the two diffusion models. Note that different DGMs have different recommended training strategies and optimal training parameters. Because the recommended default hyperparameters have been optimized for visual quality and FID scores, the defaults were used for all trainings unless specified otherwise. For a fair comparison, each DGM was trained with a fixed compute budget for a certain image size. This budget was 900 gpu-hours on a Nvidia GTX TitanX GPU for all SCMs (image size 256×256), and 900 gpu-hours on a Nvidia Quadro RTX 8000 for the VT-SOM (image size 512×512). This budget was chosen to match the compute requirements of the DGMs trained in chapter 3.

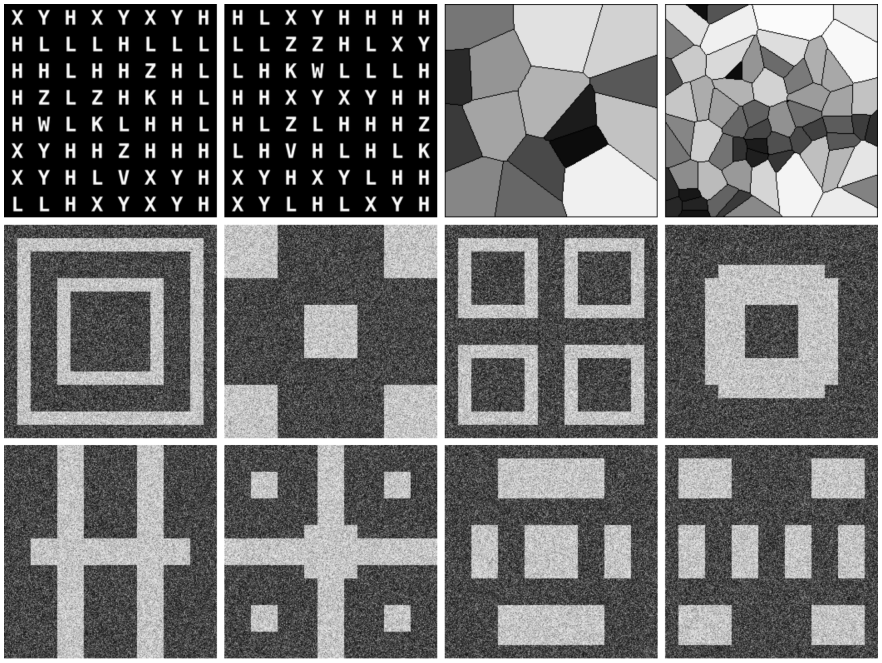


Figure 5.1: Sample realizations from all three SCMs are shown. Top row: Two realizations each from the single-class A-SCM (left) and the four-class V-SCM (right). Realizations from the V-SCM represent classes 16 and 64 respectively. Rows 2 and 3: A realization from each of the eight classes in the F-SCM.

Two variants of the DDPM: class-conditioned DDPM, and foundational DDPM were also trained/fine-tuned to assess if these variations improved the performance of DDPM. The class-conditioned DDPM learns the distribution of the data in a class-specific manner, i.e., it also takes a class label as an input during training and sampling. Class-conditioning provides control over the composition of the ensemble. The class-conditioned DDPM was trained on the two multi-class SCMs: V-SCM and F-SCM. Foundational models are typically models that are pre-trained on an extremely large dataset, and maybe employed on another dataset after finetuning/ transfer learning. The publicly available foundational DDPM, pre-trained on the ImageNet dataset [46] was fine-tuned for V-SCM, and the more complex VT-SOM dataset. In case of the VT-SOM, the foundational DDPM available for image size 512×512 was modified to bypass class conditioning so that this model could be employed for unconditional image generation. All other training parameters were retained as default for both, class-conditioned and foundational DDPMs.

The second diffusion model, LDM, consisted of training a VAE on the original images, followed by training a DDPM on the latent representation of the original images encoded

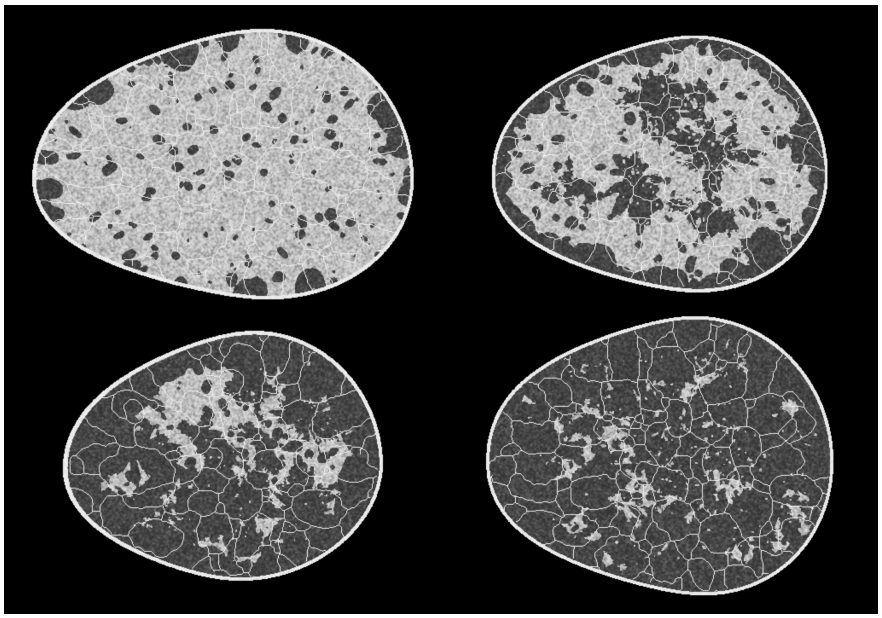


Figure 5.2: Sample images from each of the four classes in the VT-SOM. Sample realizations from (top row: L to R) dense, heterogeneous, (bottom row: L to R) scattered and fatty breast types are shown.

by the trained VAE. The default dimensions of the latent representation were retained for VT-SOM (8×128×128), and lowered to (3×64×64) for V-SCM, which consisted of simpler and smaller images as compared to VT-SOM. In all cases, the last model was chosen for analysis.

5.5 Results

It is noted that the performance of DGMs may vary with the choice of training hyper-parameters, or even random initialization. The performance reported in this work is only representative of typically trained models and may not indicate the best performance possible for any DGM; identifying the "best possible" instance of a DGM is a massive computational undertaking, and a fundamentally different problem than the goal of this work.

5.5.1 Results from the Alphabet SCM

Sample realizations from the DDPM trained on A-SCM are shown in Figure 5.3; high visual quality was observed in the generated samples, and was also represented in the near-perfect FID score of 0.1.

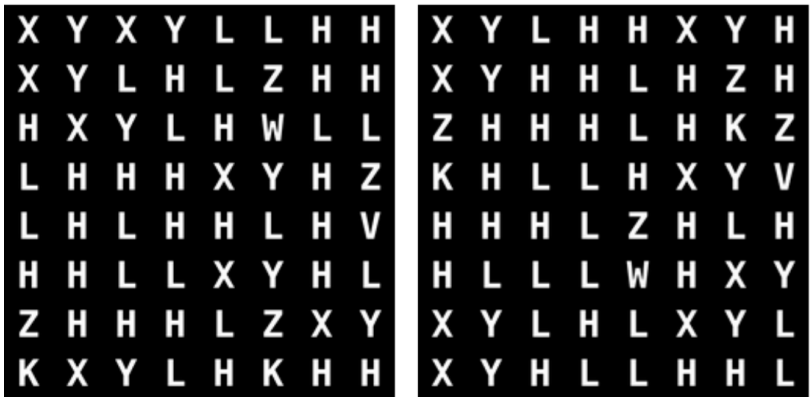


Figure 5.3: Visually high quality generated samples from DDPM (left) and SG2 (right).

As before, only realizations within which all letters were visually recognizable were included for further analysis and recognition was automated via a pattern match filter [220]. All DDPM realizations exhibited only recognizable letters, unlike both GANs employed in chapter 3. Single letter prevalence was assessed via a chi-squared goodness-of-fit test with the critical value set to 95%. About 99% of all DGMs realizations were acceptable, and 98% DDPM realizations exhibited perfect prevalence. This is in stark contrast to the GAN-generated realizations which demonstrated perfect prevalence rarely, and only by chance, as reported in chapter 3.

Results from the reproducibility of feature-pair prevalences strengthen this finding. All four letter-pairs prescribed in the training dataset were almost perfectly replicated throughout the DDPM-generated ensemble, unlike in the GAN-generated ensembles (refer Figure 3.9). Some errors in DDPM realizations are shown in Figure 5.4. These examples demonstrate that DDPM occasionally creates new pairings, or displays pairings too frequently, even when realizations are otherwise excellent. Thus, by analogy, an ensemble of DDPM-generated biomedical images could appear perfect via spot-checks, and pass traditional tests of distribution similarity, but still include images that are anatomically nonsensical.

Figure 5.4: Contextual errors were observed in some DDPM-generated realizations from A-SCM. These manifested as incorrect pairings of letters (yellow) or incorrect per-image prevalence of letter-pairs (blue). In the training data, the letter-pairs X-Y, and Z-V were always in order, and the letter-pair Z-K occurred exactly twice in each image.

5.5.2 Results from the Voronoi SCM

Samples from DGM-generated ensembles are shown in Figure 5.5. High visual similarity was observed for samples from all DGMs. The corresponding FID values were: 1.5 (DDPM), 14.1 (LDM).

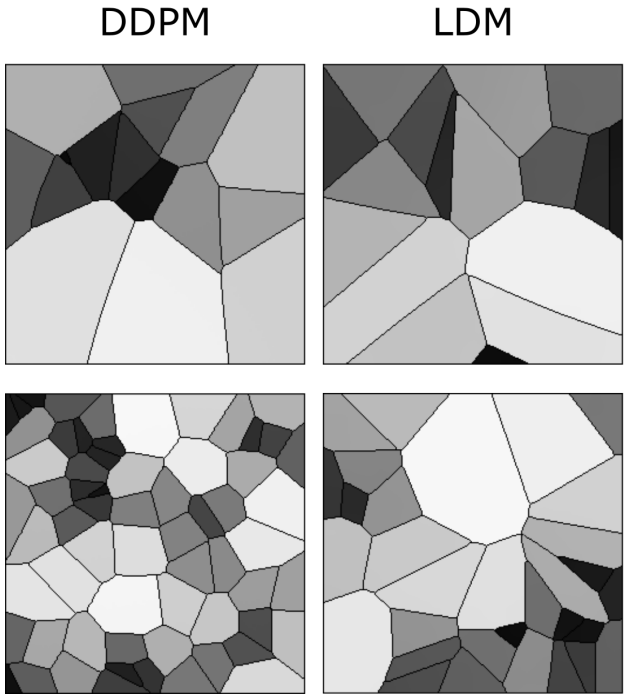


Figure 5.5: Sample DGM-generated images from the V-SCM demonstrate high visual quality for both DGMs.

Similar to the tests described in chapter 3, the DGM-generated images corresponding to the V-SCM were tested for (i) *explicit* contextual rules of shading and prevalence prescribed in the training ensemble, as well as (ii) certain *implicit* contextual features that emerge as a result of the stochastic processes defined in the SCM.

The prescribed perfect correlation between grayscale intensity and area, within each image was observed to be lower in both DGM-generated ensembles. Approximately 4% and 1% of the DDPM-, and LDM-generated ensembles respectively demonstrated a Spearman rank-correlation $\rho < 0.9$, indicating that the quantitative value of these realizations is partially lost. Next, per-image feature prevalence encoded as the number of Voronoi regions in an image was tested. Recall that, here, the number of regions defines class (refer section 3.3.1 for class prediction). It was observed that no DGM reproduced the prescribed uniform class prevalence exactly (see Figure 5.6), although both diffusion models demonstrated good mode coverage, and LDM retained the distinct modes in the training data. As reported in subsection 3.4.2, no GAN reproduced the expected class prevalence either, and one of the GANs also suffered from mode collapse. This is in accordance to the literature, where diffusion models have been reported to have good mode coverage, unlike GANs [50].

The DDPM was observed to interpolate between modes such that a substantial number of generated realizations are not any of the classes seen in the training data (see Figure 5.6), similar to GANs (shown in subsection 3.4.2). Furthermore, the DDPM unequivocally extrapolated beyond the extreme classes in the dataset (see Figure 5.8 bottom row). These observations could imply that interpolation and extrapolation are functionally equivalent.

Interpolation effects were further explored by assessing the implicit context typically arising in Voronoi diagrams [167] as before (subsection 3.4.2). Recall that the following per-image statistics were chosen to represent implicit context: number of Voronoi regions, number of junctions, junction density, mean and standard deviation of Voronoi edge lengths, mean and standard deviation of the area of a Voronoi region. Results from principal component analysis performed on these statistics are shown in Figure 5.7.

It was observed that although both diffusion models respected class-specific implicit context, DDPM-generated realizations interpolated between classes following the trend in implicit context defined by the training data. This was confirmed by visual spot checks of sample realizations from the new classes and their placement in the PCA plots. This result suggests

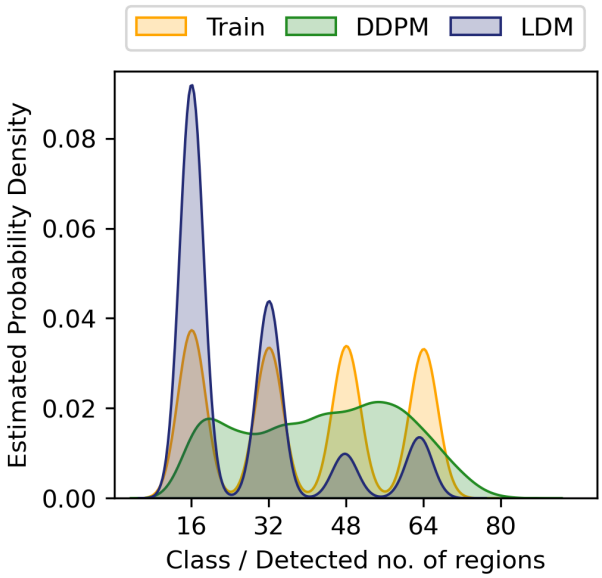


Figure 5.6: Class-prevalence results from the V-SCM demonstrated via kernel density estimates (KDE) of the data. All DGMs fail to replicate the prescribed uniform class prevalence. The DDPM demonstrates interpolation between the four distinct classes in the training dataset. In addition, DDPM also extrapolates beyond the extreme class (class 64) generating realizations corresponding to class 80, which was absent in the training dataset. Although LDM retains the distinct modes in the data, the uniform class prevalence is not respected.

that the DDPM generated a substantial number of realizations from new classes (via interpolation), but, perhaps more importantly, that those realizations may be genuine Voronoi diagrams. In section 5.6, this result is discussed further. This was in contrast to one of the GANs tested earlier: StyleGAN2, which may be more prone to errors in implicit context for a similar interpolation between classes, suggesting that at least a fraction of the interpolated SG2 images may not be considered Voronoi diagrams. The LDM demonstrated negligible interpolation between classes, and slight extrapolation of each class, suggesting that class identity was strongly captured in the latent representation of the training dataset.

Although DDPM demonstrated contextually correct class interpolation, occasional errors in statistics representing implicit context were visually observed in DDPM-generated images (see Figure 5.8), indicating that all implicit contextual features were not always perfectly reproduced. Thus, results from the V-SCM indicate that a large fraction of DDPM-generated images, but not all images, may be contextually correct in terms of quantitative meaning and class identity.

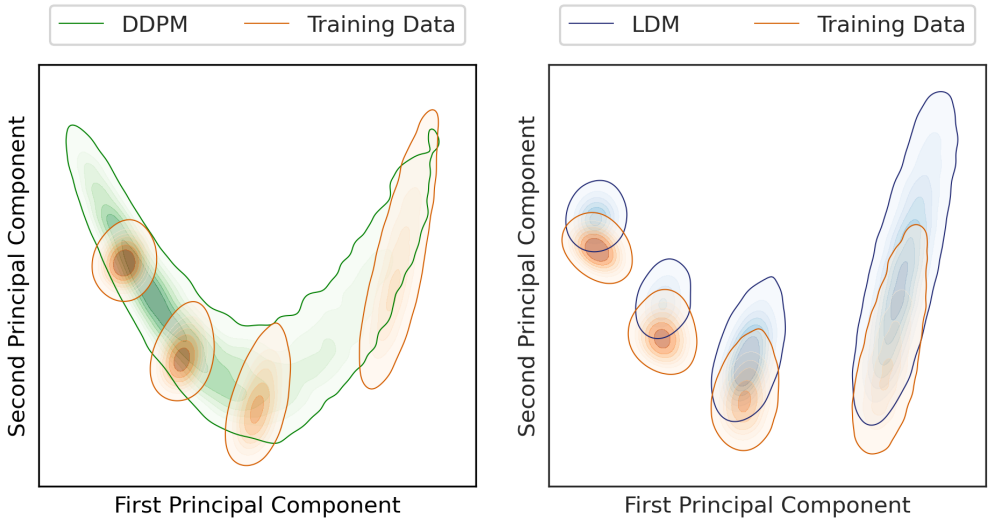


Figure 5.7: Results from the V-SCM. Principal component analysis (PCA) of the statistics representing implicit context demonstrates that interpolation between classes also resulted in an interpolation of the emergent implicit context in case of the DDPM (left). The LDM (right) generally respected the distinct classes and their respective implicit context. Note that the PCA plots are represented via kernel density estimation for display.

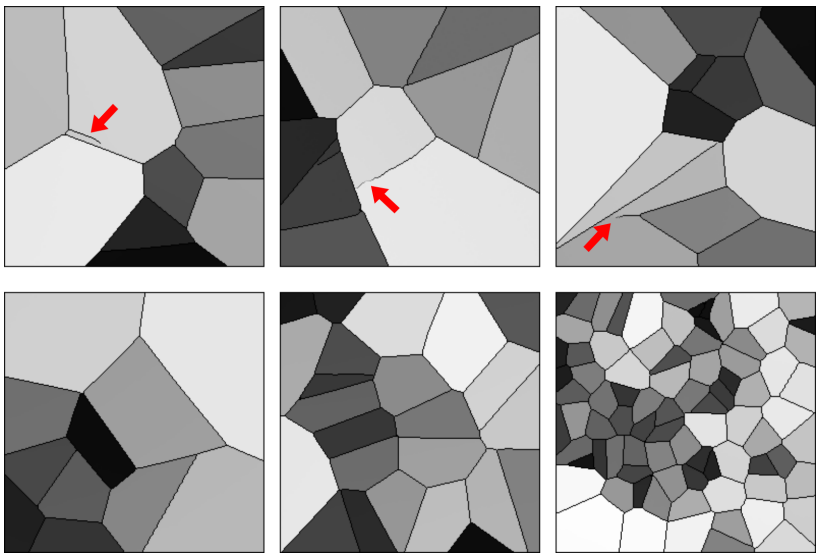


Figure 5.8: DDPM-generated samples from V-SCM exhibit implicit contextual errors like disjoint Voronoi edges (top row) and explicit errors like incorrect region count (bottom row). Although realizations in the bottom row are visually acceptable, the number of regions per-image indicating class is lower than (left), interpolated between (center), or extrapolated beyond, the classes in the training data (right)

5.5.3 Results from the Flags SCM

Sample realizations with high visual quality for the DDPM trained on the F-SCM are shown in Figure 5.9. The corresponding FID value was 5.7.

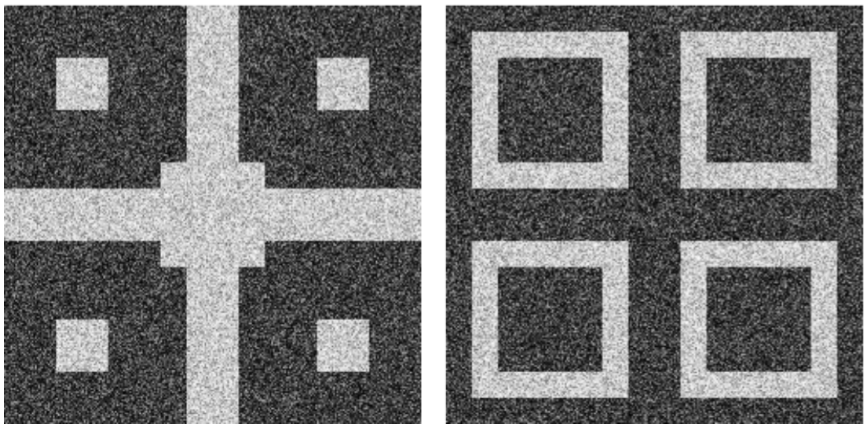


Figure 5.9: Visually high quality generated samples from the DDPM trained on the F-SCM.

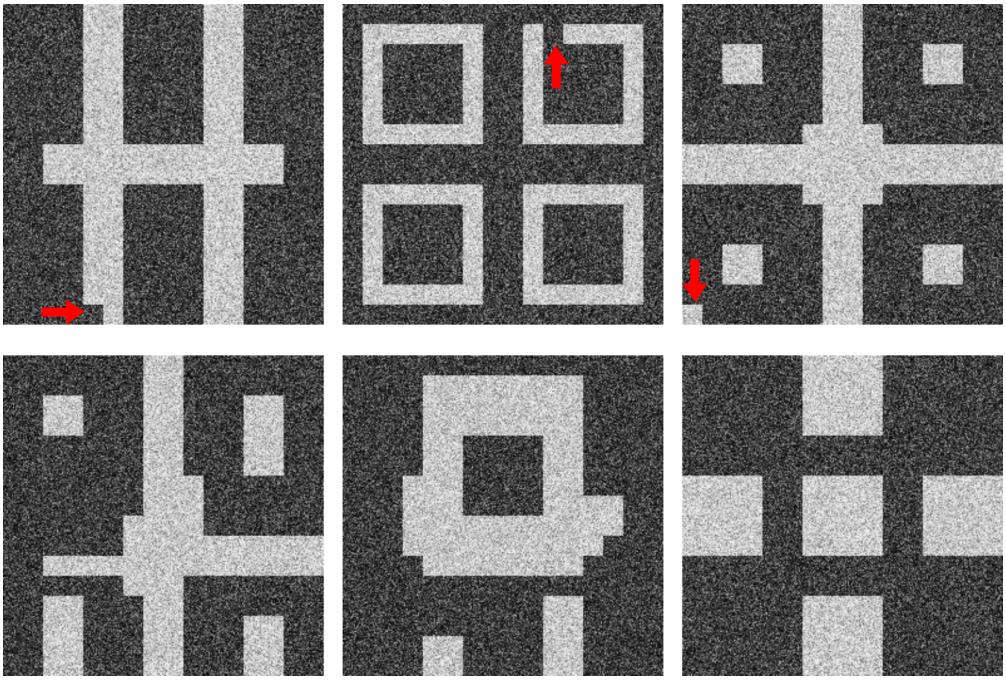


Figure 5.10: Contextually incorrect DDPM-generated samples from the F-SCM are shown. Top row: Minor errors in the foreground patterns due to a single misplaced tile were observed. Bottom row: Major errors in the class-specific foreground patterns were also observed. None of the foreground patterns in this row were present in the training data.

Table 5.2: Results from the F-SCM. Percentage of acceptable realizations in an ensemble is reported for all DGMs and for four contextual constraints. DDPM slightly outperforms SG2 in most cases. Results for prevalence and position are reported together because these constraints jointly define the foreground structure representative of a class.

Constraints	Measure of error	DDPM		SG2	
		FG	BG	FG	BG
Prevalence + Position	RMAE	99	99	98	98
Intensity	χ^2	0	0	0	9
Texture	Moran's I	100	99	96	95

Quantitative results from the F-SCM that encodes joint contextual constraints in per-image feature prevalence, position, grayscale intensity, and texture are given in Table 5.2. The class-specific foreground patterns representing joint constraints in position and prevalence were correctly reproduced by the DDPM for over 98% of the ensemble. This is also visually evident in the DDPM-generated images (see Figure 5.9).

However, errors in foreground patterns such as those shown in Figure 5.10 were observed in about 1% of the DDPM-generated ensemble. This is an important observation and the learning behavior of DDPM is discussed in detail in section 5.6. Note that the errors always occurred as misplaced or absent foreground tiles. Additionally, tiles which are never foreground in any class appeared as foreground in 0.1% of the DDPM-generated ensemble, but never in the SG2-generated ensemble (subsection 3.4.1). This suggests that the DDPM learned individual motifs that create foreground patterns instead of entire image-level patterns. Texture arising from the randomness in pixel placement was correctly reproduced in over 99% of DDPM-generated ensembles, measured as described in subsection 3.4.1. Last, the prescribed per-image intensity distributions as measured via the χ^2 goodness-of-fit test (at 95% critical value) over each image were assessed. All images from the DDPM were beyond the 99.5th percentile of the value of the χ^2 statistic computed on the training data separately for the foreground and background intensity distributions. These results might indicate some difficulty in learning multiple joint contextual constraints at once. Furthermore, the results also highlight the potential of a SCM-based evaluation approach, wherein contextual constraints are progressively added for the assessment of DGMs.

5.5.4 Results from the VICTRE SOM

Recall that the top-ranked entry in the Grand Challenge described in chapter 4 was a conditional LDM trained on the VICTRE SOM; sampling from this trained model was followed by post-hoc processing of the generated images to improve their visual quality. In this subsection, results are reported from a DDPM, which has greater model capacity (and compute requirements) than the LDM. Results from a LDM trained in-house are also provided as a baseline. The in-house LDM differs from the top-ranked entry in three ways: (i) class information was not explicitly provided to the model for training or inference, (ii) no post-hoc processing was performed on the generated image ensemble—this may result in some loss of image quality as compared to the model in chapter 4, and (iii) computational constraints of the Challenge were not applicable and longer training times were possible. Thus, the in-house trained LDM provides a fair comparison to the DDPM for the unconditional image synthesis task.

Images generated from the two diffusion models trained on the VT-SOM are shown in Figure 5.11. Images from the DDPM and LDM demonstrated high visual similarity with the training data (see Figure 5.11) as well as low (<10) FID scores; DDPM-generated images in particular, had distinctly superior visual image quality. The corresponding FID scores for the DGMs were: 1.3 (DDPM), 14.3 (LDM).

Results from the VT-SOM demonstrate that the DDPM clearly outperforms the LDM on almost all feature sets (see Table 5.3) included in the study, namely, texture features, morphology features, skeleton statistics, and the ratio of fatty to glandular tissue. (See subsection 4.4.3 for a description of the evaluation framework.)

This effect was particularly strong for morphology features (KS statistic values for DDPM, LDM: 0.049, 0.160) and skeleton statistics (KS statistic values for DDPM, LDM: 0.006, 0.109). The reproducibility of F/G ratio (also representative of class) was very similar for both models. Some DDPM-samples with extreme F/G ratio, not seen in the training, were also observed; the breast region in these samples seemed to be formed almost entirely of glandular tissue. Random samples of 200 images each from the training and DDPM-generated ensembles were visually inspected by non-domain-experts for any immediately obvious errors. Occasional artifacts in ligament structures were visually noticeable in the DDPM-generated images (see Figure 5.12). While major breaks in ligaments were observed

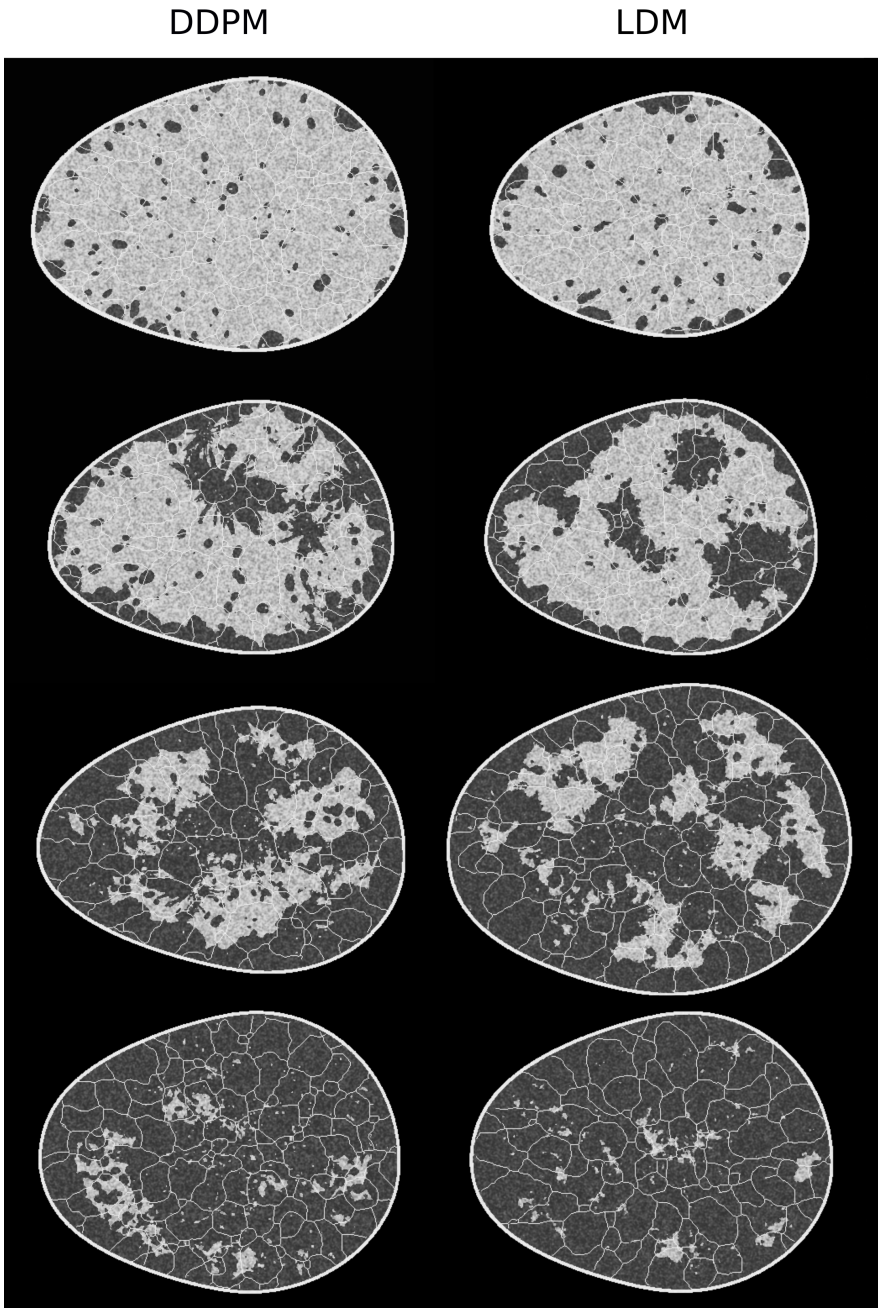


Figure 5.11: DGM-generated samples with high visual quality, corresponding to all four classes in the VT-SOM are shown. Recall that, here, the fat-to-glandular (F/G) ratio defines class.

in 1 in 6 images in the training ensemble, this rate doubled to 1 in 3 images in the DDPM-generated ensemble. These results indicate that even though DDPM outperformed all DGMs, it routinely synthesizes images with anatomical artifacts that can be spotted in the ligament

Table 5.3: Results from the VT-SOM for various feature families. Most feature families were better reproduced in the DDPM-generated ensemble as compared to the LDM-generated ensembles, as indicated by the lower KS statistic for the former.

Feature set / DGM	DDPM	LDM
	KS statistic ↓	
Texture features	**0.028**	0.085
Morphology features	**0.049**	0.129
Skeleton statistics	**0.006**	0.102
F/G ratio	0.230	**0.227**
Overall	**0.028**	0.140

Table 5.4: Analysis of class prevalence, coverage and density demonstrate the superior performance of the DDPM in representing all classes present in the training data. The class-wise prevalence in the training data was: 10%, 40%, 40%, 10%. (* indicates that skeleton statistics were excluded in this computation.)

DGM	Class prevalence (%)	Class coverage [139]	Class density [139]
DDPM	21,44,29,6	0.97, 0.96, 0.91, 0.91	0.99, 1.01, 0.98, 1.02
LDM	11,26,34,29	0.82, 0.59, 0.60, 0.49	0.98, 1.00, 0.92, 0.55

structures even by a non-domain-expert upon casual inspection. This should be taken into account before using the realizations for decision support.

Next, class-wise analysis was performed after predicting class labels on generated ensembles by employing a classifier with a VGG-16 [202] backbone as described in subsection 4.4.3. All four classes from the training ensemble were well represented in both generated ensembles (see Table 5.4). Class coverage and density [139] results demonstrate that the LDM-generated images had moderate to high class density (indicative of class fidelity) and moderate coverage (intra-class diversity) as compared to the training data. On the other hand, the DDPM-generated images demonstrated high coverage for all classes, in addition to nearly perfect class density, thus, clearly outperforming the LDM.

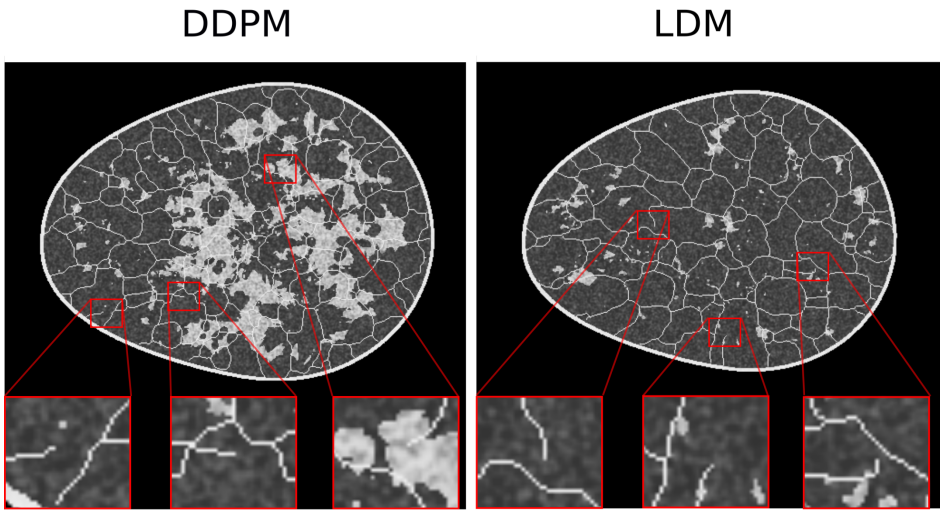


Figure 5.12: Samples from DGMs trained on VT-SOM show varied artifacts. Both, the DDPM and the LDM images exhibit strong visual quality but reveal structural errors like broken ligaments (inset).

5.5.5 Results from Variations of the DDPM

Class-Conditioned DDPM

To assess if class-conditioning can alleviate class interpolation and extrapolation observed in the DDPM, two class-conditioned DDPMs were trained on the V-SCM and the F-SCM. For both cases, the generated images were visually very similar to the unconditional DDPM images. In the first case (V-SCM), class-conditioning seemed to ensure that the distinct classes in the training data were retained well and interpolation between classes was absent. The class-specific implicit contextual features were also well replicated as observed in Figure 5.13. However, similar to the unconditional DDPM (see Figure 5.7), the exact distribution of the training data was not matched, albeit in a class-specific manner. Thus, class-conditioning may aid only in retaining distinct modes in the data as identified by labels, but not the exact class range.

In the second case (F-SCM), the foreground patterns that are indicative of class were assessed. As opposed to results from the unconditional DDPM, where forbidden foreground

regions were not respected due to class interpolation in some cases, the conditional DDPM-generated ensemble never violated this rule. This supports the finding in case of the V-SCM that labeled classes are largely respected by a class-conditioned DDPM.

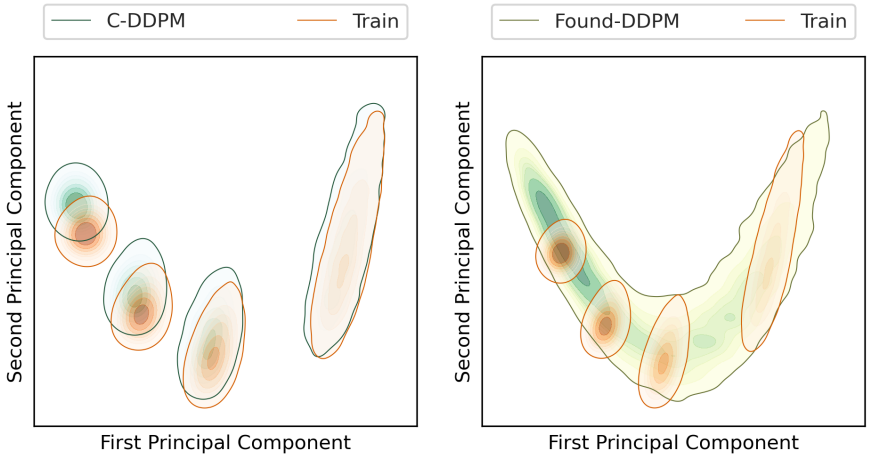


Figure 5.13: Results from the V-SCM for DDPM variants demonstrated via principal component analysis (PCA). Class-conditional DDPM (left) respects the four distinct modes in the data, but demonstrates unequal intra-class coverage and extrapolation. Foundational DDPM (right) performs very similar to the unconditional DDPM (Figure 5.7) and provides slightly better coverage for some classes. Note that the PCA plots are represented via kernel density estimation of the data for display.

Foundational DDPM

Two foundational DDPMs pre-trained on the ImageNet dataset were employed on the V-SCM and VT-SOM datasets to assess if they performed better than the generic DDPM. As foundational DDPMs have a greater model capacity than the generic DDPM and are pre-trained on a large dataset, it is possible that they may outperform generic DDPMs. This study aims to investigate the generalizability of foundational DDPMs. Recall that foundational DDPMs were afforded the same compute budget for fine-tuning as that afforded to the generic DDPM trained from scratch on our datasets.

Visual quality of the V-SCM images generated from the foundational DDPM was on par with those from the generic DDPM. For the foundational DDPM trained on V-SCM, it was observed that 9% of the realizations from the foundational DDPM demonstrated a Spearman rank-correlation $\rho < 0.9$, which represents the correlation between grayscale intensity

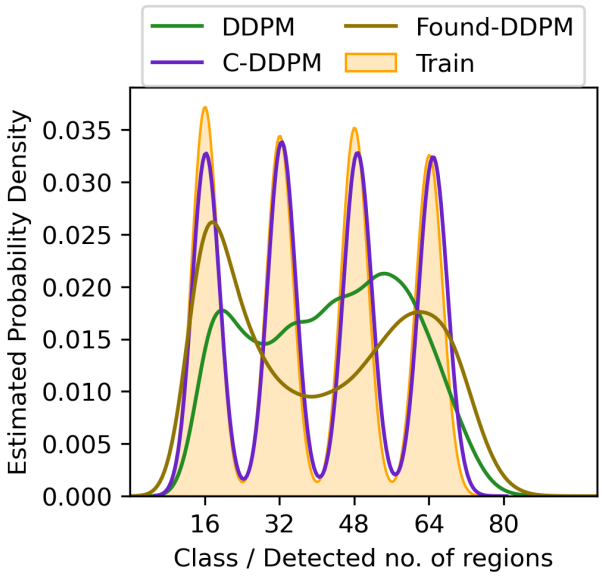


Figure 5.14: Class-prevalence results from the V-SCM for DDPM variants demonstrated via kernel density estimates. Class-conditional DDPM (C-DDPM) demonstrates an excellent match with the training data in terms of class modes and prevalence. However, foundational DDPM (foundational DDPM) demonstrated similar effects: mode coverage, class interpolation and extrapolation, as compared to the unconditional generic DDPM.

and area of Voronoi regions in an image, as compared to 4% from the unconditional DDPM ensemble. This indicates that about twice as many images in the foundational DDPM ensemble had *lower* quantitative fidelity than those in the generic DDPM ensemble. Furthermore, the reproducibility of implicit context in the DGM-generated images from the V-SCM was only slightly better for foundational DDPM as compared to the generic DDPM in terms of class coverage as seen in Figure 5.13 (right). Similar to the generic DDPM, the foundational DDPM demonstrated excellent mode coverage but did not respect the uniform class prevalence in the training data. Thus, in case of the V-SCM, both foundational DDPM and DDPM achieved similar results despite foundational DDPM having an advantage over DDPM in terms of pre-training and model capacity.

For the VT-SOM, the visual quality of images generated from the foundational DDPM was slightly inferior to those generated from the DDPM. An example each of a visually high quality image, and an unrealistic image, from the foundational DDPM ensemble are shown in Figure 5.15. This qualitative result was also reflected in the quantitative results described in Table 5.5 for all feature families and class-based analyses. However, it is noted

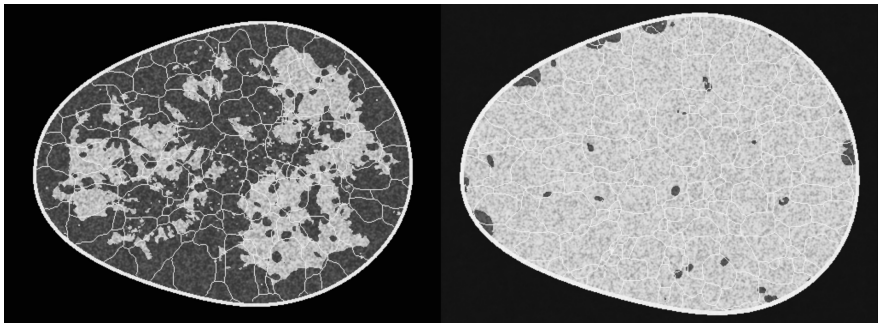


Figure 5.15: A visually good image (left), and an unrealistic image (right) in terms of ligament structure and extreme F/G ratio, generated from the foundational DDPM fine-tuned on the VT-SOM are shown. The visual image quality of these images is slightly lower than those generated from the generic DDPM.

that training the foundational DDPM beyond the specified compute budget improved its performance to approximately match the generic DDPM (results not shown). Thus, the generalization capacity of the foundational DDPM might be constrained by the similarity between the datasets employed for pre-training and fine-tuning, as expected. Hence, in some cases, a generic DDPM might provide superior performance at lower computational cost. We discuss this result further in section 5.6.

Table 5.5: Results from the VT-SOM for the foundational DDPM. In all cases, the generic DDPM outperformed foundational DDPM, given a fixed compute budget. However, it was observed that additional training of the foundational DDPM brought its performance at par with the generic DDPM (results not shown). Here, the KS statistic should ideally be 0 and the class coverage and density should be approximately 1.

Feature set / DGM	DDPM	foundational DDPM
	KS statistic ↓	
Texture features	**0.028**	0.134
Morphology features	**0.049**	0.085
Skeleton statistics	**0.006**	0.049
F/G ratio	**0.230**	0.460
Overall	**0.028**	0.069
Class prevalence(%)	21,44,29,6	41,37,20,2
Class coverage [139]	0.97, 0.96, 0.91, 0.91	0.81, 0.89, 0.64, 0.57
Class density [139]	0.99, 1.01, 0.98, 1.02	0.59, 0.98, 0.76, 1.01

5.6 Discussion

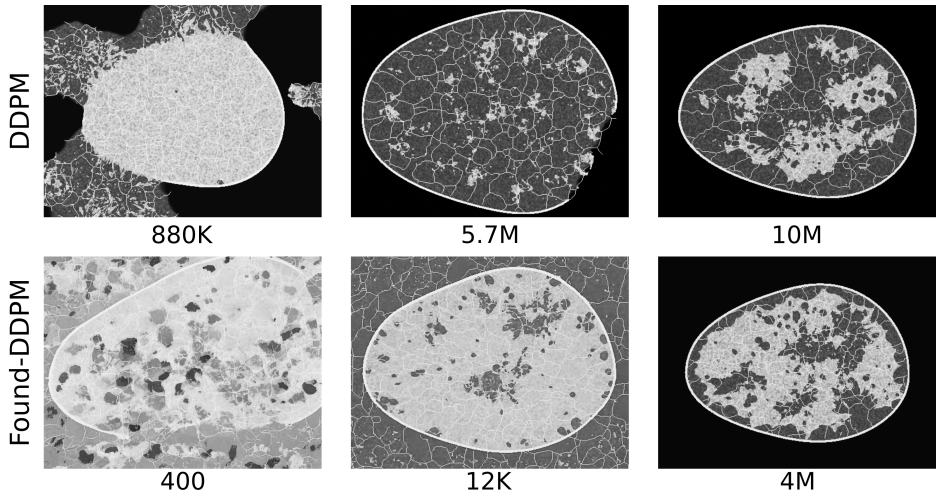


Figure 5.16: Visually interesting but random examples from the training trajectories of the DDPM and the foundational DDPM models employed on VT-SOM. The training step corresponding to each image is indicated below the image, and represents the number of images seen in training. The DDPM seemed to shape clearly demarcated zero-valued background and foreground textures, while the foundational DDPM seemed to unlearn placing textures in patches or all over the image before learning the expected features.

Over the last few years, diffusion models have emerged as a more popular alternative to GANs, especially in the absence of major computational constraints. As opposed to GANs, which implicitly model the data distribution via latent-based modeling, DDPMs estimate the data distribution by employing a theoretical foundation that is well-established in thermodynamics. Furthermore, latent representations in DDPMs typically respect the pixel coordinate system and do not decrease the dimensionality of the data, thus, potentially contributing to more reproducible long-range correlations per-image as compared to GANs. The excellent visual quality of DDPM-generated images has also contributed to their popularity. However, domain-specific errors might be present even in images that appear "perfect" to non-domain experts. The frameworks proposed in the previous chapters provide valuable tools to assess if diffusion models that produce high-quality images, also produce domain-accurate images. Although contextual errors have been known to occur in the GAN family of DGMs [137, 215, 220, 221], to our knowledge, this is the first study to demonstrate and quantify various contextual errors in a diffusion-based generative model. Results from our studies demonstrate that impactful errors likely are present in every DGM-generated ensemble. The impact of those errors is task-dependent and, therefore, should be studied case-by-case.

Even the simple SCMs employed in this work reveal stark differences in image representation within DGM paradigms. For example, differences in the characteristic image representation of the DDPM and StyleGAN2 (SG2) were observed, particularly, via the nature of the artifacts (see Figure 5.17) and training trajectories Figure 5.18. While artifacts in the DDPM generally seemed to involve misplaced but correct motifs, the artifacts in SG2 demonstrated a blending of various motifs within the same image. Similar differences were observed in the training trajectories (Figure 5.18). The DDPM seemed to first learn local elements required to construct image-level structure followed by combinations of these elements, while SG2 seemed to learn image structure through blob-like elements. Each SCM in this work represents a different context; together, the SCMs constitute a readily interpretable and intuitive method for the objective assessment of DGMs. The SCMs successively encode an increasing number of contextual constraints; this enables a step-wise evaluation of the capacity of any DGM to reproduce individual and joint contextual constraints (see Table 5.1). For example, the DDPM almost perfectly replicates the letter prevalences in the A-SCM, and largely reproduces the contextual constraints of shading and prevalence in the V-SCM, but fails to reproduce the intensity distributions in the F-SCM. This suggests that the joint replication of multiple contextual constraints remains a challenge for the DDPM.

The relevance of the evaluation approach to medical imaging applications employed in this work lies not in anatomical realism, but in the logically analogous representation of contextual attributes relevant to biomedical imaging. For example, one work [221] reported contextual errors in GAN-generated images such as misplaced pacemakers in chest radiographs. Analogously, misplaced tiles were observed in the DDPM-generated Flags-SCM images, thus exposing the capacity of DDPMs to misplace features in forbidden areas. Other examples of biomedical imaging scenarios that involve the studied contextual attributes include: (i) pathology images, wherein the cell-specific size, intensity distribution and per-image prevalence may be characteristic of different pathologies and (ii) the relative positions of organs, and the per-image prevalence of ribs in a chest radiograph. These examples are respectively analogous to: (i) the Voronoi SCM, which encodes context via shading and prevalence at multiple length scales, and (ii) the Alphabet SCM, which encodes context via per-image prevalence and relative positions of letters.

One key finding is that implicit context *for new classes* was very well reproduced in the DDPM-generated Voronoi SCM ensemble (see Figure 5.7). Because the mathematical properties of Voronoi diagrams are well established, a "ground truth" is available to test if new

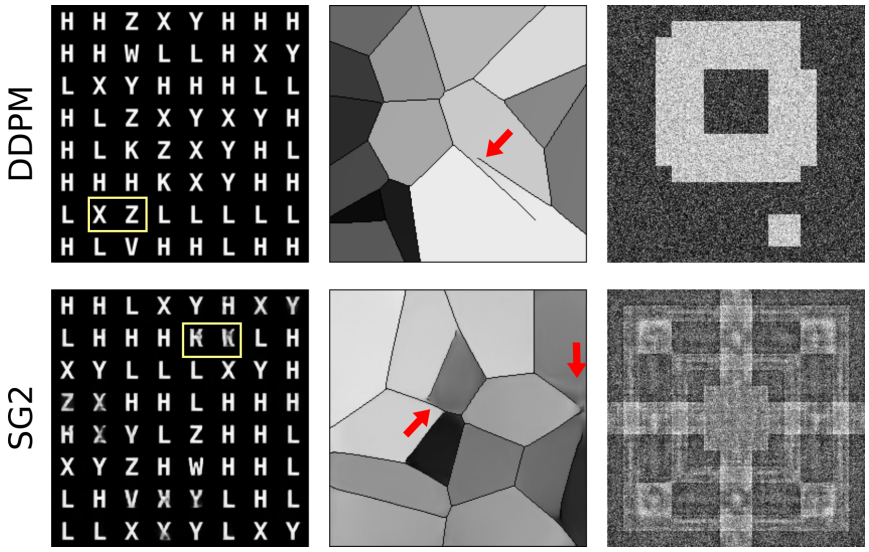


Figure 5.17: Examples of artifacts present in generated realizations from the DDPM and SG2 are shown for the three SCMs. Errors in the DDPM-generated images demonstrate misplaced but distinct motifs from the training data, whereas errors in SG2-generated images demonstrate malformations or blending of distinct motifs. This effect is seen across all SCMs.

realizations are genuine Voronoi diagrams. Thus, the finding that new classes of Voronoi also respect the expected implicit context suggests that DDPMs hold promise for data augmentation applications. With the availability of the domain-appropriate ground truth, i.e., Voronoi diagrams, additional experiments can be designed. These may include: assessing data sufficiency for learning context and experiments relating the composition of the image ensemble to learning of certain classes. Another approach to study DGMs via a surrogate ground truth is to leverage established stochastic models and adapt them to ensure recoverability of context in the generated images. This approach was demonstrated via the adapted VICTRE SOM as described in chapter 4. Adaptations to the VICTRE SOM such as prescribed intensity distributions, addition of texture, and minor processing of ligaments ensured that several contextual tests could be designed for assessing DGMs. This was possible because the context obtained in the training data was also recoverable from generated ensembles produced by reasonably well-trained DGMs.

A second result, also pertaining to interpolation across classes, was observed via the F-SCM. Specifically, interpolated instances in the DDPM-generated ensemble were not merely a linear combination of foregrounds in the training data. Furthermore, some instances also violated the regions forbidden as foreground across all classes. However, the grid in the

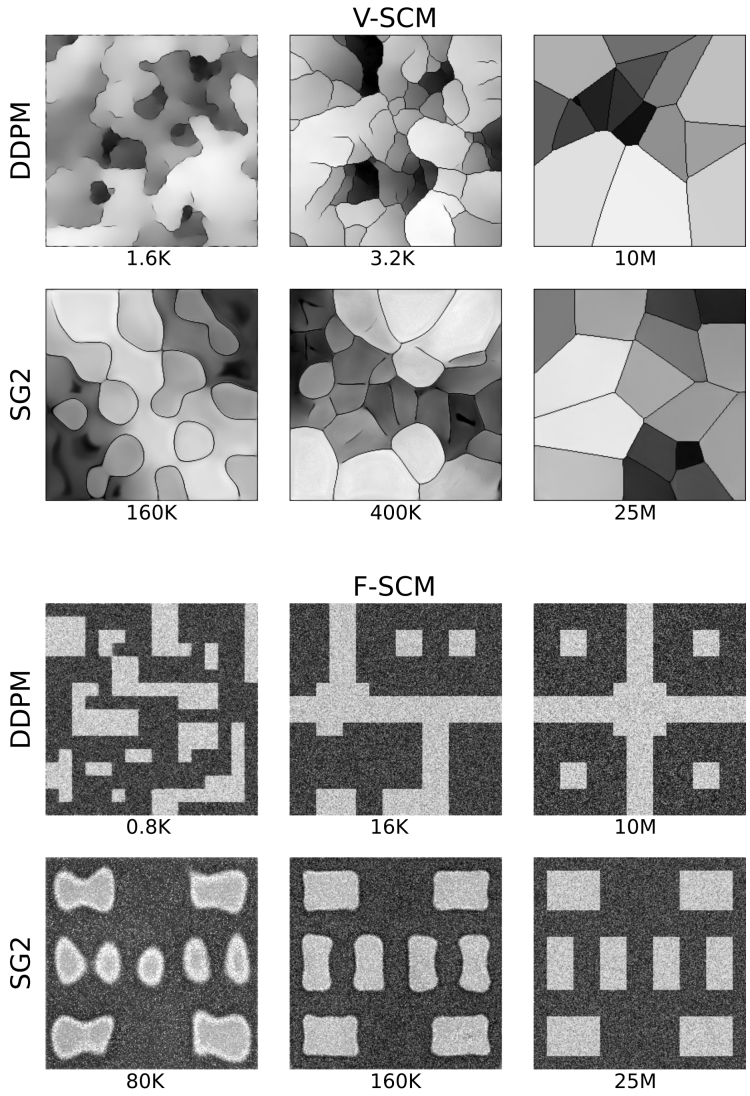


Figure 5.18: Visually interesting random examples from the training trajectories of DDPM and SG2 models employed on V-SCM and F-SCM datasets are shown. The training step corresponding to each image is indicated below the image, and represents the number of images seen in training. DDPM seemed to first learn local elements that constitute the expected structure, while SG2 seemed to learn image structure by moulding blob-like elements.

F-SCM design, or size of a single tile, was correctly learned by the DDPM, and it seems that this knowledge was employed in class interpolation. Thus, although the DDPM correctly identified the relevant local scale in the formation of patterns, it did not perfectly capture the image-level context in the F-SCM. On a more complex dataset: the VT-SOM, the DDPM largely failed to capture all contextual constraints at once. Despite the high visual quality of DDPM images, errors in ligament formation were identified even by a non-domain-expert

in about 30% of the ensemble, similar to the instances reported in chapter 4. Furthermore, unrealistic extrapolation beyond all classes in the training data was observed in the V-SCM and F-SCM, potentially due to the likelihood-based approach of DDPM that also contributes to excellent mode coverage [50]. Unlike the DDPM, the latent diffusion model employed on the V-SCM and VT-SOM, seemed to retain distinct classes. Especially, in the case of V-SCM (see Figure 5.6), all four classes were distinctly formed but classes with low number of regions were preferentially generated. It is possible that the latent encoding of the LDM captures class information, constraining not just class interpolation but potentially also intra-class diversity as observed in the results from the VT-SOM. We also explored the possibility that a class-conditioned DDPM may alleviate the issue of class interpolation. The class-conditioned DDPM retained distinct classes but demonstrated the same effects (mismatched class-specific distributions) for *each* class that an unconditional DDPM did over the entire data distribution. This implies that class-conditioning may only avoid drastic interpolation in classes as determined by the labels, but not necessarily within a class or even in attributes unrelated to class labels.

Another important factor highlighted through our results is the relation between a model's intended use and the datasets employed for training, i.e., DGM generalizability. Our experiments demonstrated that although foundational DDPMs provide a powerful alternative to the generic DDPM, they may actually perform worse than the latter in some cases. In case of the VT-SOM, although foundational DDPM and DDPM both had the same compute budget, the foundational DDPM may have spent part of training unlearning ImageNet features before learning VT-SOM features. This is consistent with the training trajectories observed in Figure 5.16. Possibly, foundational DDPM first had to learn that breast slices occupy only the central portion of an image and all textural features are contained within the egg-like shape of the slice, unlike the ImageNet where texture features may be translation invariant and the entire image is often non-zero. On the other hand, the generic DDPM seemed to learn to create a distinct zero-valued background early in training, followed by learning shape and texture at once. Thus, DGMs aimed at improving performance on natural image synthesis may not provide the same potential benefits for medical imaging tasks, and hence, domain and task relevant evaluation remains critical before deploying state-of-the-art DGMs in medical imaging workflows.

5.7 Conclusion for Chapter 5

In this chapter, an instance of the denoising diffusion probabilistic model (DDPM) was evaluated to gain insights into its capacity to reproduce contextual attributes analogous to anatomical constraints present in medical imaging scenarios. The DDPM-generated ensembles in this study demonstrated low contextual error-rates, but none of the ensembles reproduced the expected context perfectly. This evaluation goes beyond earlier evaluations of the DDPM that employed ensemble-based evaluation measures designed for natural images, or conventional measures of image quality. It is anticipated that the employed evaluation framework might yield insights into emerging DGMs and have a broader impact on decision-making and DGM benchmarking.

Chapter 6

Discussions and Conclusions

"We can only see a short distance ahead, but we can see plenty there that needs to be done." - Alan Turing

Biomedical imaging is one domain where mistakes can jeopardize patient lives. Hence, the evaluation of novel technologies is critical before their deployment in biomedical imaging workflows. This is especially true for learning-based methods. The limitations of learning-based methods, e.g., DGMs, are not obvious, unlike methods whose behavior can be sufficiently described by their mathematical formulation alone. It is possible, in case of DGMs, that overfitting or high perceptual quality of the generated data could mask hallucinations and lead to an overestimation of their general capacities. Hence, the evaluation of DGMs should go beyond the evaluation of perceptual quality by non-domain-experts and encompass domain knowledge as well.

The evaluation of DGMs cannot be relegated to only the stage immediately prior to deployment but should be a part of a feedback loop in the method design/ adaptation stage as well. Thus, several benchmarks are necessary at different stages to ensure that a DGM: possesses the technical capacity for a task, retains the diagnostic value in each image, respects the ensemble distribution, and will reliably generate an ensemble at least as effective as the training data for a downstream task and without any negative impacts.

In this thesis, the focus of DGM evaluation is on designing tests of the technical capacity of a DGM in a domain-relevant manner, with potential implications for its diagnostic applicability. These tests may be deployed in the initial stages of method development. The proposed evaluation frameworks are necessary but not sufficient tests for assessing DGMs in a manner relevant to biomedical imaging. Evaluation at a later stage in the development/ deployment workflow may involve expert reader studies, or diagnostic-task-specific evaluations designed in consultation with domain experts.

6.1 Summary and Discussions of the Major Findings from this Thesis

An important finding from this work was that none of the DGM-generated ensembles were perfect in the designed tests of spatial context. Some of these DGMs have been reported to produce extremely realistic natural images and also received very low FID scores in our studies with non-natural images. These findings highlight the fact that DGMs that perform well on natural images, or on figures of merit designed for natural images, may not necessarily be as successful in generating biomedical images, as observed in results from chapter 4. Hence, establishing domain-relevant benchmarks for biomedical images is essential and natural image benchmarks may not suffice.

Another major finding is that purposefully designed data can prove to be an effective way of probing DGM capacity relevant to biomedical imaging. The level of realism in the designed data can be varied according to the purpose of the evaluation [133]. Even with low realism in the modeled image features, the encoded spatial context provides a ground truth for testing the applicability of DGMs in a variety of scenarios relevant to biomedical imaging. In fact, low realism and high interpretability in image features, is a highly effective strategy for assessing DGMs in a *general* manner, i.e., not restricted to a single realistic imaging system/ anatomy; this is underlined by the results from the stochastic context models and their implications for biomedical imaging scenarios.

The stochastic context models designed as part of the first evaluation framework proposed in this thesis have a low level of realism and general applicability. The strength of these models lies in their interpretability, which in turn enables the creation of a known ground truth. Due to the interpretability of SCMs, it is easy for a non-clinical expert to identify potential problems arising due to the lack of DGM capacity for a certain task, even before deploying a DGM on complex, biomedical dataset. Furthermore, as visual evaluation of biomedical images requires domain expertise, it is possible that some issues identified via SCMs might have been impossible to identify for a non-expert looking at generated biomedical images.

The SCMs are not limited to the designs proposed in this thesis, and novel SCMs can also be designed according to the requirements of new tasks as long as the task-relevant contextual attribute can be encoded in a recoverable manner. A more realistic and complex

stochastic model of anatomy was adapted and employed as a part of the second evaluation framework. If a DGM could perfectly reproduce the expected context in an object, it would then be a candidate for more complex evaluations with higher level of realism and the inclusion of the effects of an imaging system. Like the studies with SCMs, the SOM-based studies also demonstrated that the kinds of errors made by many different DGMs, or artifacts induced by these DGMs, were not unique to specific DGMs. Similar artifacts were observed across multiple DGMs, implying that the current approaches to DGM design are not perfect and routinely result in specific types of artifacts. Identification of these artifacts can spur fundamental improvements in DGM approaches, and also aid the development of post-hoc processing methods for improving image quality, specific to certain DGMs before their deployment.

In the same study involving the stochastic model of anatomy, spatial context was represented via conventional image statistics. These statistics have been developed and employed over several decades; they are mathematically interpretable and can be visually related to the observed images. Although DGMs seek to represent images via a latent space, which is often not interpretable, the representation of images via conventional statistics enables a data-driven assessment of the capacities of a DGM.

A major advantage of the evaluation frameworks proposed in this thesis is that they are model-agnostic. Therefore, although the present thesis reports results from the currently popular/ state-of-the-art DGMs, novel DGMs can be benchmarked against these DGMs in future. The effects of minor architectural modifications to a DGM in a task-specific or even task-agnostic manner could be tested via the proposed methods. Thus, the stochastic models employed in this work could provide insights into the effectiveness of architectural modifications to a DGM even during the protoyping phase of model development.

Well-studied scientific problems or models provide one way of testing DGM capacity and ensuring that the generated ensemble is realistically diverse. One example was the Voronoi SCM described in chapter 3. Voronoi provides a unique solution to a space-partitioning problem and the mathematical properties of this model are well established. It was observed in a study employing the Voronoi SCM that a DGM generated contextually correct novel images in the majority of the ensemble (DDPM trained on the Voronoi SCM described in chapter 5). This is an important finding which suggests that some DGMs may hold promise

for data augmentation applications. This finding could be made primarily because a well-studied, established stochastic model was employed. In other words, established stochastic models with known properties can enable if DGMs can generate novel images, while retaining their "identity" as determined by the expected properties. Several other stochastic models exist, and can be employed similarly to the Voronoi dataset for DGM assessments. Broadly, such assessments ensure that if a DGM is to be deployed for a scientific application, it does not violate established natural laws while aiming for novelty/diversity.

6.2 Discussions and Future Work

The popularity of DGMs in biomedical imaging research has been rapidly increasing, even as errors made by DGMs are occasionally reported. This is possibly because DGMs hold tremendous potential for making biomedical imaging workflows more accurate and efficient. Broadly speaking, DGMs could be employed to overcome issues of data insufficiency, poor data quality, and missing data in practical biomedical imaging scenarios. This potential has translated to the high volume of research into DGMs for tasks such as data augmentation, image denoising or superresolution, and domain transfer [1,3,4]. Novel applications of DGMs as well as novel DGM paradigms continue to emerge. For example, a recent novel application of DGMs includes employing clinical records to generate biomedical images [225]. In parallel, DGMs with unprecedented learning capacities are also being developed, e.g., DGMs that generalize to multiple tasks [226], or learn from multi-modal data that includes text, images, and 1D-signals [227]. Thus, the use of DGMs in biomedical imaging is only expected to rise further in the coming years.

Major improvements in DGM design over the last five years have been marked via substantial improvements in visual image quality, and the capacity to generate increasingly large images at high visual quality. These improvements have often aimed to alleviate instances of poor visual image quality reported in the state-of-the-art approaches at the time, typically for natural images. When novel DGMs are adapted for use with biomedical images, a non-domain-expert may not always be able to identify if high visual quality images are anatomically unrealistic. As a result, it is possible that artifacts/ hallucinations in DGM-generated biomedical images occur more often than they are found. Several works have

reported the presence of hallucinations in various biomedical imaging modalities and for different DGMs as described in chapter 1. However, even when hallucinations are reported in biomedical images, the rate of hallucinations is rarely, if ever, reported.

A major contribution of this thesis is that certain kinds of potentially domain-relevant errors, i.e., hallucinations, can be identified, and the error-rates quantified in DGM-generated image ensembles. As the proposed methods are model agnostic, they can be employed towards benchmarking several DGMs. Furthermore, the evaluation framework based on stochastic context models, does not model specific anatomy or a specific task, but encodes general contextual attributes relevant in biomedical imaging scenarios. Thus, this test-bed generalizes to many different imaging modalities and tasks, and can be employed to rule out DGMs before they are considered for additional evaluations specific to an imaging modality/ task. The other evaluation framework that is based on the VICTRE stochastic object model provides a more complex dataset, while also enabling the quantification of DGM-specific artifacts. Together, all datasets enable a systematic evaluation of DGMs for various kinds of errors that may affect the reproducibility of domain-relevant spatial context and potentially impact decision-making.

The present work largely focuses on the evaluation of unconditional DGMs for image synthesis, and image-conditioned DGMs are only briefly explored in chapter 3. The current approach of evaluating DGMs via spatial context can also be extended to image-conditioned and text-conditioned DGMs. In case of image-conditioned DGMs, two aspects would be tested: (i) replication of context conditioned on one domain (ii) replication of context present exclusively in the output domain. In case of text-conditioned DGMs, a definition of context for text, and mapping a correspondence between spatial and text-based context would have to be established to identify the limitations of these DGMs.

Besides different kinds of DGMs, existing stochastic *object* models could also be adapted for assessing contextual correctness. The adaptation could involve additional realism, a diagnostic task, or a different imaging modality. Similarly, additional stochastic *context* models could also be designed for a given clinical task. This would require knowledge of several aspects including, but not limited to, the general contextual attribute involved in the task, the extent of the image (i.e., size in pixels) that is diagnostically relevant, the expected morphology and intensity distributions, recoverability of the relevant contextual attribute from the generated images, and a test to determine the acceptability of the generated images.

The level of realism and complexity of the SCMs can be greater than those proposed in this thesis, as long as their interpretability is retained and the recoverability of encoded features is ensured. Thus, the design of novel SCMs might involve a trade-off between interpretability and clinical realism.

Last, even after a DGM is deployed (because errors in generated images were rare), identification of the rare, unrealistic images is essential. One way to address this would be to adapt anomaly detection methods that can be deployed together with the DGM to ensure another barrier against errors cascading to downstream tasks.

6.3 Conclusion

Although DGMs are rapidly evolving for natural image applications, many novel or emerging DGMs cannot yet be employed in biomedical imaging applications. Adaptation of DGMs for biomedical imaging may require not only modifications in DGMs but also, more importantly, domain-relevant benchmarks. In the absence of domain-relevant benchmarks, if DGMs are deployed based on evaluations on natural image benchmarks alone, potentially, hallucinations could remain undetected and negatively impact patient outcomes.

Generated images can be inaccurate in different ways; hence, only a single number cannot be employed as a figure-of-merit to evaluate DGMs. A series of tests that assess different aspects of the generated images is necessary to enable a comprehensive assessment of DGMs. In this thesis, data-driven methods have been developed for assessing the capacity of DGMs to reproduce external information, that is, spatial context relevant to biomedical imaging. No modern DGM employed in the undertaken studies demonstrated perfect performance on all tests. This suggests that DGM approaches have substantial scope for improvement before being considered reliable for deployment in biomedical scenarios. At the same time, improvements in DGMs will necessitate the development of improved evaluation methods. The model-agnostic evaluation frameworks presented in this thesis are a step towards the identification of the limitations of current DGM approaches and the development of comprehensive evaluation frameworks for DGMs in general, before the potential of DGMs in biomedical imaging can be truly and safely realized.

References

[1] X. Yi, E. Walia, and P. Babyn. Generative adversarial network in medical imaging: A review. *Medical image analysis*, 58:101552, 2019.

[2] N. K. Singh and K. Raza. Medical image generation using generative adversarial networks: A review. *Health informatics: A computational perspective in healthcare*, pages 77–96, 2021.

[3] S. Kazeminia, C. Baur, A. Kuijper, B. van Ginneken, N. Navab, S. Albarqouni, and A. Mukhopadhyay. GANs for medical image analysis. *Artif. Intell. in Med.*, 109:101938, 2020.

[4] A. Kazerouni, E. K. Aghdam, M. Heidari, R. Azad, M. Fayyaz, I. Hacihaliloglu, and D. Merhof. Diffusion models in medical imaging: A comprehensive survey. *Medical Image Analysis*, page 102846, 2023.

[5] A. Kebaili, J. Lapuyade-Lahorgue, and S. Ruan. Deep learning approaches for data augmentation in medical imaging: A review. *Journal of Imaging*, 9(4):81, 2023.

[6] F. Garcea, A. Serra, F. Lamberti, and L. Morra. Data augmentation for medical imaging: A systematic literature review. *Computers in Biology and Medicine*, 152:106391, 2023.

[7] A. Kaur and G. Dong. A complete review on image denoising techniques for medical images. *Neural Processing Letters*, 55(6):7807–7850, 2023.

[8] J. M. Wolterink, T. Leiner, M. A. Viergever, and I. Išgum. Generative adversarial networks for noise reduction in low-dose ct. *IEEE transactions on medical imaging*, 36(12):2536–2545, 2017.

[9] J. Zhang, W. Gong, L. Ye, F. Wang, Z. Shangguan, and Y. Cheng. A review of deep learning methods for denoising of medical low-dose ct images. *Computers in Biology and Medicine*, page 108112, 2024.

[10] 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 *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part III 11*, pages 455–463. Springer, 2018.

[11] Y. Lei, J. Harms, T. Wang, Y. Liu, H.-K. Shu, A. B. Jani, W. J. Curran, H. Mao, T. Liu, and X. Yang. Mri-only based synthetic ct generation using dense cycle consistent generative adversarial networks. *Medical physics*, 46(8):3565–3581, 2019.

[12] D. Nie, R. Trullo, J. Lian, C. Petitjean, S. Ruan, Q. Wang, and D. Shen. Medical image synthesis with context-aware generative adversarial networks. In *Medical Image Computing and Computer Assisted Intervention- MICCAI 2017: 20th International Conference, Quebec City, QC, Canada, September 11-13, 2017, Proceedings, Part III 20*, pages 417–425. Springer, 2017.

[13] B. Bai, X. Yang, Y. Li, Y. Zhang, N. Pillar, and A. Ozcan. Deep learning-enabled virtual histological staining of biological samples. *Light: Science & Applications*, 12(1):57, 2023.

[14] H.-M. Zhang and B. Dong. A review on deep learning in medical image reconstruction. *Journal of the Operations Research Society of China*, 8:311–340, 2020.

[15] Y. Song, L. Shen, L. Xing, and S. Ermon. Solving inverse problems in medical imaging with score-based generative models. *arXiv preprint arXiv:2111.08005*, 2021.

[16] J. Xuan, Y. Yang, Z. Yang, D. He, and L. Wang. On the anomalous generalization of gans. *arXiv preprint arXiv:1909.12638*, 2019.

[17] S. U. Dar, M. Yurt, L. Karacan, A. Erdem, E. Erdem, and T. Cukur. Image synthesis in multi-contrast mri with conditional generative adversarial networks. *IEEE transactions on medical imaging*, 38(10):2375–2388, 2019.

[18] C. K. Chong and E. T. W. Ho. Synthesis of 3d mri brain images with shape and texture generative adversarial deep neural networks. *IEEE Access*, 9:64747–64760, 2021.

[19] M. F. Ng and C. A. Hargreaves. Generative adversarial networks for the synthesis of chest x-ray images. *Engineering Proceedings*, 31(1):84, 2023.

[20] K. Packhäuser, L. Folle, F. Thamm, and A. Maier. Generation of anonymous chest radiographs using latent diffusion models for training thoracic abnormality classification systems. In *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)*, pages 1–5. IEEE, 2023.

[21] J. Lee, T. Mustafaev, and R. M. Nishikawa. Impact of gan artifacts for simulating mammograms on identifying mammographically occult cancer. *Journal of Medical Imaging*, 10(5):054503–054503, 2023.

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

[23] X. Yi and P. Babyn. Sharpness-aware low-dose ct denoising using conditional generative adversarial network. *Journal of digital imaging*, 31:655–669, 2018.

[24] Y. Wang, B. Yu, L. Wang, C. Zu, D. S. Lalush, W. Lin, X. Wu, J. Zhou, D. Shen, and L. Zhou. 3d conditional generative adversarial networks for high-quality pet image estimation at low dose. *Neuroimage*, 174:550–562, 2018.

[25] D. Mahapatra and B. Bozorgtabar. Retinal vasculature segmentation using local saliency maps and generative adversarial networks for image super resolution. *arXiv preprint arXiv:1710.04783*, 2017.

[26] J. M. Wolterink, A. M. Dinkla, M. H. Savenije, P. R. Seevinck, C. A. van den Berg, and I. Išgum. Deep mr to ct synthesis using unpaired data. In *Simulation and Synthesis in Medical Imaging: Second International Workshop, SASHIMI 2017, Held in Conjunction with MICCAI 2017, Québec City, QC, Canada, September 10, 2017, Proceedings 2*, pages 14–23. Springer, 2017.

[27] S. Bhadra, W. Zhou, and M. A. Anastasio. Medical image reconstruction with image-adaptive priors learned by use of generative adversarial networks. In *Medical Imaging 2020: Physics of Medical Imaging*, volume 11312, pages 206–213. SPIE, 2020.

[28] J. Liu, R. Anirudh, J. J. Thiagarajan, S. He, K. A. Mohan, U. S. Kamilov, and H. Kim. Dolce: A model-based probabilistic diffusion framework for limited-angle ct reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10498–10508, 2023.

[29] J. Wolleb, F. Bieder, R. Sandkühler, and P. C. Cattin. Diffusion models for medical anomaly detection. volume 13438, pages 35–45. Springer, 2022.

[30] W. H. L. Pinaya, M. S. Graham, R. Gray, P. F. Da Costa, P.-D. Tudosiu, P. Wright, Y. H. Mah, A. D. MacKinnon, J. T. Teo, R. Jager, D. Werring, G. Rees, P. Nachev, S. Ourselin, and M. J. Cardoso. Fast Unsupervised Brain Anomaly Detection and Segmentation with Diffusion Models. *arXiv:2206.03461*, 2022.

[31] P. Costa, A. Galdran, M. I. Meyer, M. Niemeijer, M. Abràmoff, A. M. Mendonça, and A. Campilho. End-to-end adversarial retinal image synthesis. *IEEE transactions on medical imaging*, 37(3):781–791, 2017.

[32] Q. Wang, Y. Chen, N. Zhang, and Y. Gu. Medical image inpainting with edge and structure priors. *Measurement*, 185:110027, 2021.

[33] K. Armanious, V. Kumar, S. Abdulatif, T. Hepp, S. Gatidis, and B. Yang. ipa-medgan: Inpainting of arbitrary regions in medical imaging. In *2020 IEEE international conference on image processing (ICIP)*, pages 3005–3009. IEEE, 2020.

[34] G. Müller-Franzes, J. M. Niehues, F. Khader, S. T. Arasteh, C. Haarburger, C. Kuhl, T. Wang, T. Han, T. Nolte, S. Nebelung, et al. A multimodal comparison of latent denoising diffusion probabilistic models and generative adversarial networks for medical image synthesis. *Scientific Reports*, 13(1):12098, 2023.

[35] J. P. Cohen, M. Luck, and S. Honari. Distribution matching losses can hallucinate features in medical image translation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part I*, pages 529–536. Springer, 2018.

[36] R. Durall, M. Keuper, and J. Keuper. Watch your up-convolution: Cnn based generative deep neural networks are failing to reproduce spectral distributions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7890–7899, 2020.

[37] K. Schwarz, Y. Liao, and A. Geiger. On the frequency bias of generative models. *Advances in Neural Information Processing Systems*, 34:18126–18136, 2021.

[38] M. Khayatkhoei and A. Elgammal. Spatial frequency bias in convolutional generative adversarial networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 7152–7159, 2022.

[39] V. Cheplygina. Cats or CAT scans: Transfer learning from natural or medical image source data sets? *Current Opinion in Biomedical Engineering*, 9:21–27, 2019.

[40] Y. Chen, X.-H. Yang, Z. Wei, A. A. Heidari, N. Zheng, Z. Li, H. Chen, H. Hu, Q. Zhou, and Q. Guan. Generative adversarial networks in medical image augmentation: A review. *Computers in Biology and Medicine*, 144:105382, 2022.

[41] V. Sorin, Y. Barash, E. Konen, and E. Klang. Creating artificial images for radiology applications using generative adversarial networks (GANs)–a systematic review. *Academic radiology*, 27(8):1175–1185, 2020.

[42] 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 *2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018)*, pages 240–244. IEEE, 2018.

[43] T. Rädsch, A. Reinke, V. Weru, M. D. Tizabi, N. Schreck, A. E. Kavur, B. Pekdemir, T. Roß, A. Kopp-Schneider, and L. Maier-Hein. Labelling instructions matter in biomedical image analysis. *Nature Machine Intelligence*, 5(3):273–283, 2023.

[44] A. Borji. Pros and cons of GAN evaluation measures. *Comput. Vis. Image Underst.*, 179:41–65, 2019.

[45] 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. *Advances in Neural Information Processing Systems*, 30, 2017.

[46] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[47] A. Alaa, B. Van Breugel, E. S. Saveliev, and M. van der Schaar. How faithful is your synthetic data? sample-level metrics for evaluating and auditing generative models. In *International Conference on Machine Learning*, pages 290–306. PMLR, 2022.

[48] A. Badano, C. G. Graff, A. Badal, D. Sharma, R. Zeng, F. W. Samuelson, S. J. Glick, and K. J. Myers. Evaluation of digital breast tomosynthesis as replacement of full-field digital mammography using an in silico imaging trial. *JAMA network open*, 1(7):e185474–e185474, 2018.

[49] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen. Improved techniques for training gans. *Advances in neural information processing systems*, 29, 2016.

[50] Z. Xiao, K. Kreis, and A. Vahdat. Tackling the generative learning trilemma with denoising diffusion gans. In *International Conference on Learning Representations*, 2022.

[51] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[52] P. Dhariwal and A. Nichol. Diffusion models beat GANs on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

[53] B. Julesz. Visual pattern discrimination. *IRE transactions on Information Theory*, 8(2):84–92, 1962.

[54] S. A. Roach. The theory of random clumping. *(No Title)*, 1968.

[55] M. Bertalmio, G. Sapiro, V. Caselles, and C. Ballester. Image inpainting. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, pages 417–424, 2000.

[56] A. A. Efros and T. K. Leung. Texture synthesis by non-parametric sampling. In *Proceedings of the seventh IEEE international conference on computer vision*, volume 2, pages 1033–1038. IEEE, 1999.

[57] S. C. Zhu, Y. Wu, and D. Mumford. Filters, random fields and maximum entropy (frame): Towards a unified theory for texture modeling. *International Journal of Computer Vision*, 27:107–126, 1998.

[58] J. Portilla and E. P. Simoncelli. Texture modeling and synthesis using joint statistics of complex wavelet coefficients. In *IEEE workshop on statistical and computational theories of vision*, 1999.

[59] G. Peyré. Texture synthesis with grouplets. *IEEE transactions on pattern analysis and machine intelligence*, 32(4):733–746, 2009.

[60] L. Gatys, A. S. Ecker, and M. Bethge. Texture synthesis using convolutional neural networks. *Advances in neural information processing systems*, 28, 2015.

[61] A. Oliva and A. Torralba. The role of context in object recognition. *Trends in cognitive sciences*, 11(12):520–527, 2007.

[62] S. Bond-Taylor, A. Leach, Y. Long, and C. G. Willcocks. Deep generative modelling: A comparative review of VAEs, GANs, normalizing flows, energy-based and autoregressive models. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2021.

[63] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[64] I. Kobyzev, S. Prince, and M. A. Brubaker. Normalizing flows: Introduction and ideas. *stat*, 1050:25, 2019.

[65] J. Ehrhardt and M. Wilms. Autoencoders and variational autoencoders in medical image analysis. In *Biomedical Image Synthesis and Simulation*, pages 129–162. Elsevier, 2022.

[66] F.-A. Croitoru, V. Hondru, R. T. Ionescu, and M. Shah. Diffusion models in vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[67] M. Arjovsky and L. Bottou. Towards principled methods for training generative adversarial networks. In *International Conference on Learning Representations*, 2016.

[68] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein generative adversarial networks. In *International conference on machine learning*, pages 214–223. PMLR, 2017.

[69] X. Mao, Q. Li, H. Xie, R. Y. Lau, Z. Wang, and S. Paul Smolley. Least squares generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2794–2802, 2017.

[70] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida. Spectral normalization for generative adversarial networks. In *International Conference on Learning Representations*, 2018.

[71] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville. Improved training of wasserstein gans. *Advances in neural information processing systems*, 30, 2017.

[72] A. Sauer, K. Schwarz, and A. Geiger. Stylegan-xl: Scaling stylegan to large diverse datasets. In *ACM SIGGRAPH 2022 conference proceedings*, pages 1–10, 2022.

[73] A. Brock, J. Donahue, and K. Simonyan. Large scale gan training for high fidelity natural image synthesis. In *International Conference on Learning Representations*, 2018.

[74] A. Odena, C. Olah, and J. Shlens. Conditional image synthesis with auxiliary classifier gans. In *International conference on machine learning*, pages 2642–2651. PMLR, 2017.

[75] A. Alotaibi. Deep generative adversarial networks for image-to-image translation: A review. *Symmetry*, 12(10):1705, 2020.

[76] Y. Burda, R. Grosse, and R. Salakhutdinov. Importance weighted autoencoders. *arXiv preprint arXiv:1509.00519*, 2015.

[77] M. D. Hoffman and M. J. Johnson. Elbo surgery: yet another way to carve up the variational evidence lower bound. In *Workshop in Advances in Approximate Bayesian Inference, NIPS*, volume 1, 2016.

[78] R. Child. Very deep vaes generalize autoregressive models and can outperform them on images. In *International Conference on Learning Representations*, 2020.

[79] J. Yu, X. Li, J. Y. Koh, H. Zhang, R. Pang, J. Qin, A. Ku, Y. Xu, J. Baldridge, and Y. Wu. Vector-quantized image modeling with improved vqgan. In *International Conference on Learning Representations*, 2021.

[80] Y. Song and S. Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019.

[81] A. Q. Nichol and P. Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, pages 8162–8171. PMLR, 2021.

[82] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

[83] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[84] M. Gong, S. Xie, W. Wei, M. Grundmann, K. Batmanghelich, T. Hou, et al. Semi-implicit denoising diffusion models (siddms). *Advances in Neural Information Processing Systems*, 36, 2024.

[85] L. Dinh, D. Krueger, and Y. Bengio. Nice: Non-linear independent components estimation. *arXiv preprint arXiv:1410.8516*, 2014.

[86] L. Dinh, J. Sohl-Dickstein, and S. Bengio. Density estimation using real nvp. In *International Conference on Learning Representations*, 2016.

[87] G. Papamakarios, T. Pavlakou, and I. Murray. Masked autoregressive flow for density estimation. *Advances in neural information processing systems*, 30, 2017.

[88] R. Van Den Berg, L. Hasenclever, J. M. Tomczak, and M. Welling. Sylvester normalizing flows for variational inference. In *34th Conference on Uncertainty in Artificial Intelligence 2018, UAI 2018*, pages 393–402. Association For Uncertainty in Artificial Intelligence (AUAI), 2018.

[89] W. Grathwohl, R. T. Chen, J. Bettencourt, I. Sutskever, and D. Duvenaud. Ffjord: Free-form continuous dynamics for scalable reversible generative models. In *International Conference on Learning Representations*, 2018.

[90] G. Papamakarios, E. Nalisnick, D. J. Rezende, S. Mohamed, and B. Lakshminarayanan. Normalizing flows for probabilistic modeling and inference. *The Journal of Machine Learning Research*, 22(1):2617–2680, 2021.

[91] C. Lu, J. Chen, C. Li, Q. Wang, and J. Zhu. Implicit normalizing flows. In *International Conference on Learning Representations*, 2020.

[92] H. Uzunova, M. Wilms, N. D. Forkert, H. Handels, and J. Ehrhardt. A systematic comparison of generative models for medical images. *International Journal of Computer Assisted Radiology and Surgery*, 17(7):1213–1224, 2022.

[93] A. Borji. Pros and cons of gan evaluation measures: New developments. *Computer Vision and Image Understanding*, 215:103329, 2022.

[94] Q. Xu, G. Huang, Y. Yuan, C. Guo, Y. Sun, F. Wu, and K. Weinberger. An empirical study on evaluation metrics of generative adversarial networks. *arXiv preprint arXiv:1806.07755*, 2018.

[95] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

[96] G. Parmar, R. Zhang, and J.-Y. Zhu. On aliased resizing and surprising subtleties in gan evaluation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11410–11420, 2022.

[97] A. Mathiasen and F. Hvilshøj. Backpropagating through fr\'echet inception distance. *arXiv preprint arXiv:2009.14075*, 2020.

[98] C. Nash, J. Menick, S. Dieleman, and P. Battaglia. Generating images with sparse representations. In *International Conference on Machine Learning*, pages 7958–7968. PMLR, 2021.

[99] M. J. Chong and D. Forsyth. Effectively unbiased fid and inception score and where to find them. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6070–6079, 2020.

[100] S. Jung and M. Keuper. Internalized biases in fréchet inception distance. In *NeurIPS 2021 Workshop on Distribution Shifts: Connecting Methods and Applications*, 2021.

[101] A. J. Smola, A. Gretton, and K. Borgwardt. Maximum mean discrepancy. In *13th international conference, ICONIP*, pages 3–6, 2006.

[102] L. O'Bray, M. Horn, B. Rieck, and K. Borgwardt. Evaluation metrics for graph generative models: Problems, pitfalls, and practical solutions. *arXiv preprint arXiv:2106.01098*, 2021.

[103] M. Bińkowski, D. J. Sutherland, M. Arbel, and A. Gretton. Demystifying mmd gans. In *International Conference on Learning Representations*, 2018.

[104] H. H. Barrett and K. J. Myers. *Foundations of image science*. John Wiley & Sons, 2013.

[105] Z. Ren, X. Y. Stella, and D. Whitney. Controllable medical image generation via gan. *Journal of perceptual imaging*, 5:0005021, 2022.

[106] Z. Liu, S. Wolfe, Z. Yu, R. Laforest, J. C. Mhlanga, T. J. Fraum, M. Itani, F. Dehdashti, B. A. Siegel, and A. K. Jha. Observer-study-based approaches to quantitatively evaluate the realism of synthetic medical images. *Physics in Medicine & Biology*, 68(7):074001, 2023.

[107] V. A. Kelkar, D. S. Gotsis, R. Deshpande, F. J. Brooks, K. Prabhat, K. J. Myers, R. Zeng, and M. A. Anastasio. Evaluating generative stochastic image models using task-based image quality measures. In *Medical Imaging 2023: Image Perception, Observer Performance, and Technology Assessment*, volume 12467, pages 304–310. SPIE, 2023.

[108] T. Dzanic, K. Shah, and F. Witherden. Fourier spectrum discrepancies in deep network generated images. *Advances in neural information processing systems*, 33:3022–3032, 2020.

[109] J. Frank, T. Eisenhofer, L. Schönherr, A. Fischer, D. Kolossa, and T. Holz. Leveraging frequency analysis for deep fake image recognition. In *International conference on machine learning*, pages 3247–3258. PMLR, 2020.

[110] Y. Jeong, D. Kim, S. Min, S. Joe, Y. Gwon, and J. Choi. Bihpf: Bilateral high-pass filters for robust deepfake detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 48–57, 2022.

[111] Y. Jeong, D. Kim, Y. Ro, and J. Choi. Frepgan: robust deepfake detection using frequency-level perturbations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 1060–1068, 2022.

[112] O. Giudice, L. Guarnera, and S. Battiato. Fighting deepfakes by detecting gan dct anomalies. *Journal of Imaging*, 7(8):128, 2021.

[113] A. M. Eskicioglu and P. S. Fisher. Image quality measures and their performance. *IEEE Transactions on communications*, 43(12):2959–2965, 1995.

[114] D. M. Chandler. Seven challenges in image quality assessment: past, present, and future research. *International Scholarly Research Notices*, 2013, 2013.

[115] P. F. Michael and H.-J. Yoon. Survey of image denoising methods for medical image classification. In *Medical Imaging 2020: Computer-Aided Diagnosis*, volume 11314, pages 892–899. SPIE, 2020.

[116] Z. Yu, M. A. Rahman, R. Laforest, T. H. Schindler, R. J. Gropler, R. L. Wahl, B. A. Siegel, and A. K. Jha. Need for objective task-based evaluation of deep learning-based denoising methods: a study in the context of myocardial perfusion spect. *Medical physics*, 50(7):4122–4137, 2023.

[117] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila. Analyzing and improving the image quality of styleGAN. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 8110–8119, 2020.

[118] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018.

[119] T. Karras, T. Aila, S. Laine, and J. Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. *International Conference on Learning Representations*, 2018.

[120] S. Shan, W. Ding, J. Passananti, H. Zheng, and B. Y. Zhao. Prompt-specific poisoning attacks on text-to-image generative models. *arXiv preprint arXiv:2310.13828*, 2023.

[121] P. Esser, R. Rombach, and B. Ommer. A note on data biases in generative models. *arXiv preprint arXiv:2012.02516*, 2020.

[122] A. Tsitsulin, M. Munkhoeva, D. Mottin, P. Karras, A. Bronstein, I. Oseledets, and E. Müller. The shape of data: Intrinsic distance for data distributions. In *International Conference on Learning Representations*, 2020.

[123] D. Horak, S. Yu, and G. Salimi-Khorshidi. Topology distance: A topology-based approach for evaluating generative adversarial networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 7721–7728, 2021.

[124] V. Khrulkov and I. Oseledets. Geometry score: A method for comparing generative adversarial networks. In *International conference on machine learning*, pages 2621–2629. PMLR, 2018.

[125] J. Fragemann, L. Ardizzone, J. Egger, and J. Kleesiek. Review of disentanglement approaches for medical applications–towards solving the gordian knot of generative models in healthcare. *arXiv preprint arXiv:2203.11132*, 2022.

[126] S. O'Brien, M. Groh, and A. Dubey. Evaluating generative adversarial networks on explicitly parameterized distributions. *arXiv preprint arXiv:1812.10782*, 2018.

[127] F. Regol, A. Kroon, and M. Coates. Evaluation of categorical generative models-bridging the gap between real and synthetic data. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[128] M. Lucic, K. Kurach, M. Michalski, S. Gelly, and O. Bousquet. Are gans created equal? a large-scale study. *Advances in neural information processing systems*, 31, 2018.

[129] A. Srivastava, L. Valkov, C. Russell, M. U. Gutmann, and C. Sutton. Veegan: Reducing mode collapse in gans using implicit variational learning. *Advances in neural information processing systems*, 30, 2017.

[130] M. Okawa, E. S. Lubana, R. Dick, and H. Tanaka. Compositional abilities emerge multiplicatively: Exploring diffusion models on a synthetic task. *Advances in Neural Information Processing Systems*, 36, 2024.

[131] A. Moskvichev, V. V. Odouard, and M. Mitchell. The conceptarc benchmark: Evaluating understanding and generalization in the arc domain. *Trans. Mach. Learn. Res.*, 2023.

[132] F. O. Bochud, C. K. Abbey, and M. P. Eckstein. Statistical texture synthesis of mammographic images with clustered lumpy backgrounds. *Optics express*, 4(1):33–43, 1999.

[133] A. Badano. "How much realism is needed?"—the wrong question in silico imagers have been asking. *Med. Phys.*, 44(5):1607–1609, 2017.

[134] D. Sharma, C. G. Graff, A. Badal, R. Zeng, P. Sawant, A. Sengupta, E. Dahal, and A. Badano. In silico imaging tools from the victre clinical trial. *Medical physics*, 46(9):3924–3928, 2019.

[135] E. Abadi, W. P. Segars, B. M. Tsui, P. E. Kinahan, N. Bottenus, A. F. Frangi, A. Maidment, J. Lo, and E. Samei. Virtual clinical trials in medical imaging: a review. *Journal of Medical Imaging*, 7(4):042805–042805, 2020.

[136] B. Barufaldi, A. D. Maidment, M. Dustler, R. Axelsson, H. Tomic, S. Zackrisson, A. Tingberg, and P. R. Bakic. Virtual clinical trials in medical imaging system evaluation and optimisation. *Radiation Protection Dosimetry*, 195(3-4):363–371, 2021.

[137] V. A. Kelkar, D. S. Gotsis, F. J. Brooks, K. Prabhat, K. J. Myers, R. Zeng, and M. A. Anastasio. Assessing the ability of generative adversarial networks to learn canonical medical image statistics. *IEEE Trans. Med. Imaging*, 2023.

[138] M. S. Sajjadi, O. Bachem, M. Lucic, O. Bousquet, and S. Gelly. Assessing generative models via precision and recall. *Advances in neural information processing systems*, 31, 2018.

[139] M. F. Naeem, S. J. Oh, Y. Uh, Y. Choi, and J. Yoo. Reliable fidelity and diversity metrics for generative models. In *International Conference on Machine Learning*, pages 7176–7185. PMLR, 2020.

[140] L. Simon, R. Webster, and J. Rabin. Revisiting precision and recall definition for generative model evaluation. In *International Conference on Machine Learning (ICML)*, 2019.

[141] D. Bau, J.-Y. Zhu, J. Wulff, W. Peebles, H. Strobelt, B. Zhou, and A. Torralba. Seeing what a gan cannot generate. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4502–4511, 2019.

[142] W. Xia, Y. Zhang, Y. Yang, J.-H. Xue, B. Zhou, and M.-H. Yang. Gan inversion: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3):3121–3138, 2022.

[143] J. Choi, S. Kim, Y. Jeong, Y. Gwon, and S. Yoon. Ilvr: Conditioning method for denoising diffusion probabilistic models. *arXiv preprint arXiv:2108.02938*, 2021.

[144] G. Zhou, Y. Fan, J. Shi, Y. Lu, and J. Shen. Conditional generative adversarial networks for domain transfer: A survey. *Applied Sciences*, 12(16):8350, 2022.

[145] K. Shmelkov, C. Schmid, and K. Alahari. How good is my gan? In *Proceedings of the European conference on computer vision (ECCV)*, pages 213–229, 2018.

[146] S. Santurkar, L. Schmidt, and A. Madry. A classification-based study of covariate shift in gan distributions. In *International Conference on Machine Learning*, pages 4480–4489. PMLR, 2018.

[147] L. Maier-Hein, A. Reinke, P. Godau, M. D. Tizabi, F. Buettner, E. Christodoulou, B. Glocker, F. Isensee, J. Kleesiek, M. Kozubek, et al. Metrics reloaded: recommendations for image analysis validation. *Nature methods*, pages 1–18, 2024.

[148] Z. Liu, J. C. Mhlanga, H. Xia, B. A. Siegel, and A. K. Jha. Need for objective task-based evaluation of image segmentation algorithms for quantitative pet: A study with acrin 6668/rtog 0235 multicenter clinical trial data. *Journal of Nuclear Medicine*, 65(3):485–492, 2024.

[149] D. Scholz, B. Wiestler, D. Rueckert, and M. J. Menten. Metrics to quantify global consistency in synthetic medical images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 25–34. Springer, 2023.

[150] V. A. Kelkar, X. Zhang, J. Granstedt, H. Li, and M. A. Anastasio. Task-based evaluation of deep image super-resolution in medical imaging. In *Medical Imaging 2021: Image Perception, Observer Performance, and Technology Assessment*, volume 11599, pages 207–213. SPIE, 2021.

[151] W. Zhou, S. Bhadra, F. J. Brooks, H. Li, and M. A. Anastasio. Learning stochastic object models from medical imaging measurements by use of advanced ambient generative adversarial networks. *Journal of Medical Imaging*, 9(1):015503–015503, 2022.

[152] A. Tversky. Features of similarity. *Psychological review*, 84(4):327, 1977.

[153] L. Quinn, K. Tryposkiadis, J. Deeks, H. C. De Vet, S. Mallett, L. B. Mokkink, Y. Takwoingi, S. Taylor-Phillips, and A. Sitch. Interobserver variability studies in diagnostic imaging: a methodological systematic review. *The British Journal of Radiology*, 96(1148):20220972, 2023.

[154] A. J. Barnett, F. R. Schwartz, C. Tao, C. Chen, Y. Ren, J. Y. Lo, and C. Rudin. A case-based interpretable deep learning model for classification of mass lesions in digital mammography. *Nature Machine Intelligence*, 3(12):1061–1070, 2021.

[155] W. N. von Sinner. New diagnostic signs in hydatid disease; radiography, ultrasound, ct and mri correlated to pathology. *European journal of radiology*, 12(2):150–159, 1991.

[156] D. A. Forsyth and J. Ponce. A modern approach. *Computer vision: a modern approach*, 17:21–48, 2003.

[157] R. S. Ledley and L. B. Lusted. Reasoning foundations of medical diagnosis: symbolic logic, probability, and value theory aid our understanding of how physicians reason. *Science*, 130(3366):9–21, 1959.

[158] R. Seising. From vagueness in medical thought to the foundations of fuzzy reasoning in medical diagnosis. *Artificial Intelligence in Medicine*, 38(3):237–256, 2006.

[159] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth. Describing objects by their attributes. In *2009 IEEE conference on computer vision and pattern recognition*, pages 1778–1785. IEEE, 2009.

[160] C. Chen, O. Li, D. Tao, A. Barnett, C. Rudin, and J. K. Su. This looks like that: deep learning for interpretable image recognition. *Advances in neural information processing systems*, 32, 2019.

[161] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang. Generative image inpainting with contextual attention. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5505–5514, 2018.

[162] H. Li, G. Li, L. Lin, H. Yu, and Y. Yu. Context-aware semantic inpainting. *IEEE transactions on cybernetics*, 49(12):4398–4411, 2018.

[163] W. Zhang, J. Zhu, Y. Tai, Y. Wang, W. Chu, B. Ni, C. Wang, and X. Yang. Context-aware image inpainting with learned semantic priors. *arXiv preprint arXiv:2106.07220*, 2021.

[164] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2536–2544, 2016.

[165] M. J. Fenske, E. Aminoff, N. Gronau, and M. Bar. Top-down facilitation of visual object recognition: object-based and context-based contributions. *Progress in brain research*, 155:3–21, 2006.

[166] E. Barenholtz. Quantifying the role of context in visual object recognition. *Visual Cognition*, 22(1):30–56, 2014.

[167] B. Boots, K. Sugihara, S. N. Chiu, and A. Okabe. *Spatial tessellations: concepts and applications of Voronoi diagrams*. John Wiley & Sons, 2009.

[168] P. A. Moran. Notes on continuous stochastic phenomena. *Biometrika*, 37(1/2):17–23, 1950.

[169] R. Deshpande, M. A. Anastasio, and F. J. Brooks. Evaluating the capacity of deep generative models to reproduce measurable high-order spatial arrangements in diagnostic images. In *SPIE Medical Imaging*, volume 12032, pages 521–526. SPIE, 2022.

[170] J. Nunez-Iglesias, A. J. Blanch, O. Looker, M. W. Dixon, and L. Tilley. A new python library to analyse skeleton images confirms malaria parasite remodelling of the red blood cell membrane skeleton. *PeerJ*, 6:e4312, 2018.

[171] C. Doersch, A. Gupta, and A. A. Efros. Context as supervisory signal: Discovering objects with predictable context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part III 13*, pages 362–377. Springer, 2014.

[172] S. Kaji and S. Kida. Overview of image-to-image translation by use of deep neural networks: denoising, super-resolution, modality conversion, and reconstruction in medical imaging. *Radiological physics and technology*, 12:235–248, 2019.

[173] M.-C. Yeh, S. Tang, A. Bhattad, C. Zou, and D. Forsyth. Improving style transfer with calibrated metrics. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, March 2020.

[174] M. Wright and B. Ommer. ArtFID: Quantitative evaluation of neural style transfer. In *Pattern Recognition: 44th DAGM German Conference, DAGM GCPR 2022, Konstanz, Germany, September 27–30, 2022, Proceedings*, pages 560–576. Springer, 2022.

[175] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.

[176] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.

[177] R. M. Haralick, K. Shanmugam, and I. H. Dinstein. Textural features for image classification. *IEEE Transactions on systems, man, and cybernetics*, (6):610–621, 1973.

[178] N. Bayramoglu, M. Kaakinen, L. Eklund, and J. Heikkila. Towards virtual H&E staining of hyperspectral lung histology images using conditional generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 64–71, 2017.

[179] P. Shamsolmoali, M. Zareapoor, E. Granger, H. Zhou, R. Wang, M. E. Celebi, and J. Yang. Image synthesis with adversarial networks: A comprehensive survey and case studies. *Inf. Fusion*, 2021.

[180] A. Borji. Pros and cons of GAN evaluation measures: New developments. *Comput. Vis. Image Underst.*, 2022.

[181] S. M. Astley, W. Chen, K. J. Myers, and R. M. Nishikawa. Special section guest editorial: Evaluation methodologies for clinical AI. *J. Med. Imaging*, 7(1), 2020.

[182] C. Nash, J. Menick, S. Dieleman, and P. Battaglia. Generating images with sparse representations. In *International Conference on Machine Learning*, volume 139, pages 7958–7968. PMLR, 2021.

[183] S. Zhao, H. Ren, A. Yuan, J. Song, N. Goodman, and S. Ermon. Bias and generalization in deep generative models: An empirical study. *Advances in Neural Information Processing Systems*, 31, 2018.

[184] S. R. Dolly, Y. Lou, M. A. Anastasio, and H. Li. Learning-based stochastic object models for characterizing anatomical variations. *Physics in Medicine & Biology*, 63(6):065004, 2018.

[185] A. Badano, A. Badal, S. Glick, C. G. Graff, F. Samuelson, D. Sharma, and R. Zeng. In silico imaging clinical trials for regulatory evaluation: initial considerations for victre, a demonstration study. In *Medical Imaging 2017: Physics of Medical Imaging*, volume 10132, pages 494–499. SPIE, 2017.

[186] A. Badano, M. Lago, E. Sizikova, J. Delfino, S. Guan, M. Anastasio, and B. Sahiner. The stochastic digital human is now enrolling for in silico imaging trials—methods and tools for generating digital cohorts. *Progress in Biomedical Engineering*, 5(4):042002, 2023.

[187] L. Tronchin, R. Sicilia, E. Cordelli, S. Ramella, and P. Soda. Evaluating gans in medical imaging. In *Deep Generative Models, and Data Augmentation, Labelling, and Imperfections: First Workshop, DGM4MICCAI 2021, and First Workshop, DALI 2021, Held in Conjunction with MICCAI 2021, Strasbourg, France, October 1, 2021, Proceedings 1*, pages 112–121. Springer, 2021.

[188] Y. Jang, J. Yoo, and H. Hong. Assessment and analysis of fidelity and diversity for gan-based medical image generative model. *Journal of the Korea Computer Graphics Society*, 28(2):11–19, 2022.

[189] S. Van der Walt, J. L. Schönberger, J. Nunez-Iglesias, F. Boulogne, J. D. Warner, N. Yager, E. Gouillart, and T. Yu. scikit-image: image processing in python. *PeerJ*, 2:e453, 2014.

[190] F. Li, U. Villa, S. Park, and M. A. Anastasio. 3-d stochastic numerical breast phantoms for enabling virtual imaging trials of ultrasound computed tomography. *IEEE transactions on ultrasonics, ferroelectrics, and frequency control*, 69(1):135–146, 2021.

[191] L. Liberman and J. H. Menell. Breast imaging reporting and data system (bi-rads). *Radiologic Clinics*, 40(3):409–430, 2002.

[192] D. Gotsis, V. Kelkar, R. Deshpande, F. Brooks, P. KC, K. Myers, R. Zeng, and M. Anastasio. Data for the 2023 aapm grand challenge on deep generative modeling for learning medical image statistics, https://doi.org/10.13012/B2IDB-2773204_V3, 2023.

[193] C. Castella, K. Kinkel, F. Descombes, M. P. Eckstein, P.-E. Sottas, F. R. Verdun, and F. O. Bochud. Mammographic texture synthesis: second-generation clustered lumpy backgrounds using a genetic algorithm. *Optics express*, 16(11):7595–7607, 2008.

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

[195] T. Smith Jr, G. Lange, and W. B. Marks. Fractal methods and results in cellular morphology—dimensions, lacunarity and multifractals. *Journal of neuroscience methods*, 69(2):123–136, 1996.

[196] D. Freedman and P. Diaconis. On the histogram as a density estimator: L 2 theory. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 57(4):453–476, 1981.

[197] B. B. Chaudhuri and N. Sarkar. Texture segmentation using fractal dimension. *IEEE Transactions on pattern analysis and machine intelligence*, 17(1):72–77, 1995.

[198] R. J. Prokop and A. P. Reeves. A survey of moment-based techniques for unoccluded object representation and recognition. *CVGIP: Graphical Models and Image Processing*, 54(5):438–460, 1992.

[199] J. Rogowska. Overview and fundamentals of medical image segmentation. *Handbook of medical imaging, processing and analysis*, pages 69–85, 2000.

[200] J. Flusser, B. Zitova, and T. Suk. *Moments and moment invariants in pattern recognition*. John Wiley & Sons, 2009.

[201] I. M. Chakravarti, R. G. Laha, and J. Roy. *Handbook of methods of applied statistics*. Wiley, 1967.

[202] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[203] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. 2022 ieee. In *CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10674–10685, 2021.

[204] G. Müller-Franzes, J. M. Niehues, F. Khader, S. T. Arasteh, C. Haarburger, C. Kuhl, T. Wang, T. Han, S. Nebelung, J. N. Kather, et al. Diffusion probabilistic models beat gans on medical images. *arXiv preprint arXiv:2212.07501*, 2022.

[205] Z. Xiao, K. Kreis, and A. Vahdat. Tackling the generative learning trilemma with denoising diffusion gans. *arXiv preprint arXiv:2112.07804*, 2021.

[206] T. Karras, M. Aittala, J. Hellsten, S. Laine, J. Lehtinen, and T. Aila. Training generative adversarial networks with limited data. *Advances in neural information processing systems*, 33:12104–12114, 2020.

[207] N. Wu, K. J. Geras, Y. Shen, J. Su, S. G. Kim, E. Kim, S. Wolfson, L. Moy, and K. Cho. Breast density classification with deep convolutional neural networks. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6682–6686. IEEE, 2018.

[208] M. A. Wirth. Shape analysis and measurement. *Image Processing Group*, pages 1–49, 2004.

[209] G. Bohling. Introduction to geostatistics and variogram analysis. *Kansas geological survey*, 1(10):1–20, 2005.

[210] Y. Song, L. Shen, L. Xing, and S. Ermon. Solving inverse problems in medical imaging with score-based generative models. In *ICLR*, 2022.

[211] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR, 2015.

[212] F. Khader, G. Müller-Franzes, S. Tayebi Arasteh, T. Han, C. Haarburger, M. Schulze-Hagen, P. Schad, S. Engelhardt, B. Baeßler, S. Foersch, et al. Denoising diffusion probabilistic models for 3d medical image generation. *Scientific Reports*, 13(1):7303, 2023.

[213] Z. Dorjsembe, S. Odonchimed, and F. Xiao. Three-dimensional medical image synthesis with denoising diffusion probabilistic models. In *Medical Imaging with Deep Learning*, 2022.

[214] W. H. L. Pinaya, P.-D. Tudosiu, J. Dafflon, P. F. da Costa, V. Fernandez, P. Nachev, S. Ourselin, and M. J. Cardoso. Brain imaging generation with latent diffusion models. *arXiv:2209.07162*, 2022.

[215] G. Müller-Franzes, J. M. Niehues, F. Khader, S. T. Arasteh, C. Haarburger, C. Kuhl, T. Wang, T. Han, S. Nebelung, J. N. Kather, et al. Diffusion probabilistic models beat gans on medical images. *arXiv preprint arXiv:2212.07501*, 2022.

[216] P. N. Huy and T. M. Quan. Denoising diffusion medical models. *arXiv preprint arXiv:2304.09383*, 2023.

[217] M. Iskandar, H. Mannering, Z. Sun, J. Matthew, H. Kerdegari, L. Peralta, and M. Xochicale. Towards realistic ultrasound fetal brain imaging synthesis. *arXiv preprint arXiv:2304.03941*, 2023.

[218] P. A. Moghadam, S. Van Dalen, K. C. Martin, J. Lennerz, S. Yip, H. Farahani, and A. Bashashati. A morphology focused diffusion probabilistic model for synthesis of histopathology images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2000–2009, 2023.

[219] M. Woodland, M. A. Taie, J. A. M. Silva, M. Eltaher, F. Mohn, A. Shieh, A. Castelo, S. Kundu, J. P. Yung, A. B. Patel, et al. Importance of feature extraction in the calculation of fr\'echet distance for medical imaging. *arXiv preprint arXiv:2311.13717*, 2023.

[220] R. Deshpande, M. A. Anastasio, and F. J. Brooks. A method for evaluating deep generative models of images via assessing the reproduction of high-order spatial context. *arXiv preprint arXiv:2111.12577v2*, 2023.

[221] A. DuMont Schütte, J. Hetzel, S. Gatidis, T. Hepp, B. Dietz, S. Bauer, and P. Schwab. Overcoming barriers to data sharing with medical image generation: a comprehensive evaluation. *NPJ digital medicine*, 4(1):141, 2021.

[222] A. Jalal, M. Arvinte, G. Daras, E. Price, A. G. Dimakis, and J. Tamir. Robust compressed sensing mri with deep generative priors. In *Adv Neural Inf Process Syst*, volume 34, pages 14938–14954, 2021.

[223] H. Chung and J. C. Ye. Score-based diffusion models for accelerated mri. *Med Image Anal*, 80:102479, 2022.

[224] J. Liu, R. Anirudh, J. J. Thiagarajan, S. He, K. A. Mohan, U. S. Kamilov, and H. Kim. Dolce: A model-based probabilistic diffusion framework for limited-angle ct reconstruction. *arXiv preprint arXiv:2211.12340*, 2022.

[225] T. Kikuchi, S. Hanaoka, T. Nakao, T. Takenaga, Y. Nomura, H. Mori, and T. Yoshikawa. Synthesis of hybrid data consisting of chest radiographs and tabular clinical records using dual generative models for covid-19 positive cases. *Journal of Imaging Informatics in Medicine*, pages 1–11, 2024.

[226] P. Chambon, C. Bluethgen, J.-B. Delbrouck, R. Van der Sluijs, M. Połacin, J. M. Z. Chaves, T. M. Abraham, S. Purohit, C. P. Langlotz, and A. Chaudhari. Roentgen: vision-language foundation model for chest x-ray generation. *arXiv preprint arXiv:2211.12737*, 2022.

[227] B. Azad, R. Azad, S. Eskandari, A. Bozorgpour, A. Kazerouni, I. Rekik, and D. Merhof. Foundational models in medical imaging: A comprehensive survey and future vision. *arXiv preprint arXiv:2310.18689*, 2023.

Citations Added November 7, 2024

Chapter 3:
Deshpande, R., Anastasio, M. A., & Brooks, F. J. (2024). A method for evaluating deep generative models of images for hallucinations in high-order spatial context. Pattern Recognition Letters, 186, 23–29. https://doi.org/10.1016/j.patrec.2024.08.023. Licensed under CC BY 4.0.

Chapter 4:
Deshpande R, Kelkar VA, Gotsis D, et al. (2024). Report on the AAPM grand challenge on deep generative modeling for learning medical image statistics. Med Phys., 1-17. https://doi.org/10.1002/mp.17473. Licensed under CC BY-NC-ND 4.0.

Chapter 5:
R. Deshpande, M. Özbey, H. Li, M. A. Anastasio and F. J. Brooks. (2024). Assessing the Capacity of a Denoising Diffusion Probabilistic Model to Reproduce Spatial Context, IEEE Transactions on Medical Imaging, 43:10, 3608-3620, Oct. 2024, https://doi.org/10.1109/TMI.2024.3414931. Licensed under CC BY 4.0.