When do GANs replicate? On the choice of dataset size

Qianli Feng[1,2] Chenqi Guo[1] Fabian Benitez-Quiroz[1] Aleix Martinez[1,2]
[1]The Ohio State University [2]Amazon
{feng.559, guo.1648, benitez-quiroz.1, martinez.158}@osu.edu

Abstract

Do GANs replicate training images? Previous studies have shown that GANs do not seem to replicate training data without significant change in the training procedure. This leads to a series of research on the exact condition needed for GANs to overfit to the training data. Although a number of factors has been theoretically or empirically identified, the effect of dataset size and complexity on GANs replication is still unknown. With empirical evidence from BigGAN and StyleGAN2, on datasets CelebA, Flower and LSUN-bedroom, we show that dataset size and its complexity play an important role in GANs replication and perceptual quality of the generated images. We further quantify this relationship, discovering that replication percentage decays exponentially with respect to dataset size and complexity, with a shared decaying factor across GAN-dataset combinations. Meanwhile, the perceptual image quality follows a U-shape trend w.r.t dataset size. This finding leads to a practical tool for one-shot estimation on minimal dataset size to prevent GAN replication which can be used to guide datasets construction and selection.

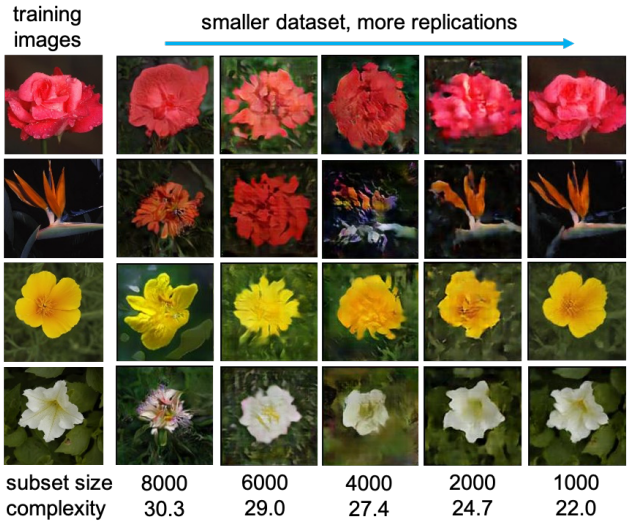

Figure 1: An example of GAN replications in Flower dataset. The complexity is measured by Intrinsic Dimensionality. We study the condition of GAN replication with a particular focus on the effect dataset size/complexity. For a given image synthesis task, when the dataset size decreases, the replication become more prominent.

1. Introduction

Generative Adversarial Networks (GANs) has attracted consistent attention since its first proposal in [9]. Since then, the photorealism of the synthetic images has seen dramatic improvement with methods like BigGAN [2], StyleGAN [12, 13], etc. This low cost generation of photo-realistic samples brings new possibility to applications like content creation and dataset augmentation, all of which are based on the assumption that the generator does not merely replicate training data. The GAN replication (or memorization, overfitting [4, 3]) problem, besides its obvious theoretical value, is also practically important.

A number of recent studies have explored the option of using GANs to augment training datasets to improve the performance of downstream machine learning algorithms [5, 25, 24, 6], especially for medical applications where patients data is scarce. When these data augmentation GANs overfit to the training data, the augmented samples will provide little or no additional value to the downstream algorithms, defying the purpose of augmenting the data. Replicating training samples is also problematic for content creation due to potential copyright infringement [7]. This is more problematic when the GAN based face swapping technique is used for preserving patient privacy (de-identification) as used in [30]. If a replication happens during the swapping process, the technique is then protecting one's privacy by costing another's portrait right and potentially spreading mis-information about the individual. Without deep understanding on the mechanism and conditions

Qianli Feng and Chenqi Guo contribute equally to the paper. Code is available at https://github.com/chenqiguo/GAN_replication. Authors were supported by NIH grant R01DC014498, R01EY020834 and HFSP grant RGP0036/2016.

of GANs replication, one might even found the issue more complicated as the replication itself might not be an intentional action from the user or the creator of the GAN. Thus, it is highly important to further our understanding on the replication behavior of GANs.

Fortunately, researchers have already started to look into the underlying mechanism and contributing factors of potential GANs replication/memorization [26, 19, 17, 1], which we discuss in Section 2. These works, although providing important insight on the effect of factors like latent code, discriminator complexity, have not yet explored the role of dataset size and complexity in GANs replication.

In this paper, we attempt to fill this gap by empirically study the relationship between GANs replication and dataset size/complexity. We show that it is possible for GANs to replicate with unmodified training procedure, challenging the common view that GANs tend not to memorize training data under a normal training setup [26]. This finding not only sheds new light on the potential mechanism of GAN replication, but also provides practical guidance on estimating minimal dataset size when new a dataset is being constructed for image synthesis purpose.

2. Related Works

Our study is not the first to study GAN replication. Studies proposing novel GAN architectures (e.g. DCGAN [19], PGAN [11]) usually show that their results are not generated by merely memorizing training samples. Repeated non-memorizing results from novel GANs architectures seem to indicate that with a normal training procedure, GANs memorization is not likely [26]. More specifically, [19] validate this information by traversing the latent space and checking for visualization and sharp changes. [26] studies GANs memorization specifically with latent code recovery and conclude that memorization is not detectable for GANs trained without cycle-consistency loss. [17] provides theoretical context for GANs memorization which further shows that the size of the support of the latent space might lead to unseen latent codes replicating training data. [1] studies the relationship between GAN memorization and discriminator size, concluding that the distributions learnt by GANs have significantly less support than real distributions. [28] then shows that with fixed latent codes during the training process the authors can achieve GAN memorization.

On the application front, [23] studied identity leakage in the face generation application with GANs. Although the topic is different from ours, the study nevertheless shows significant ethical implication on this issue. Studies like [26, 19, 17, 1] aim to address the class imbalance problem by using GANs, due to the random up/down sampling can lead to data replication. However, this assumption of non-replicating behavior of GANs is challenged by our study, leading to more cautious practices when augmenting a small dataset to address the imbalance problem.

In the previous studies investigating GAN memorization, regardless of active latent code seeking methods, or fixed latent codes, the methods deviate from training procedures originally proposed by each GAN, affecting the external validity of the conclusions. In this study, we achieve GAN replication without modifying training procedures. On the other hand, the previous studies on GAN replication have not yet explore the role of dataset size/compleity. In this study, we reveal a relationship between dataset size, GAN replication and generated image photorealism, providing a more specific guidance on deciding dataset size for image synthesis tasks.

3. Problem Definition

The specific problem concerning this study is the relationship between dataset size/complexity and GAN replication. To formulate the problem mathematically, let us denote a set of training images as $\mathcal{X}$ with each element $\mathbf{X} \in \mathbb{R}^m$. p_{data} is an empirical data distribution defined on $\mathcal{X}$ (which is usually a uniform discrete distribution over all training images). A generator $G : \mathbb{R}^k \rightarrow \mathbb{R}^m$ maps k-dimensional latent code random variables $z \sim p_{\text{latent}}$ to synthetic images $G(z)$. Then, the probability that the generator G replicates at level α with respect to metric $d(\cdot, \cdot)$, denoted as $P_\alpha(G, d, p_{\text{data}})$, is,

$$P_\alpha(G, d, \mathcal{X}) = \Pr\left(\left[\min_{\mathbf{X} \in \mathcal{X}} d(G(z), \mathbf{X})\right] \leq \alpha\right). \quad (1)$$

Ideally, we want to know the exact function $P_\alpha(G, d, \mathcal{X})$. But this is impractical as the function defined on a joint domain of all GANs, metrics and datasets is too complex. As described in Section 1, we wish to study when does a GAN replicate in terms of dataset size and complexity. Thus, in this study, we limit G and $d(\cdot, \cdot)$ while empirically study the effect of dataset on $P_\alpha(G, d, \mathcal{X})$. Let μ be a set function characterizing the training dataset $\mathcal{X}$, given a generator G and a distance metric $d(\cdot, \cdot)$, the specific function we are to study is,

$$f : \mu(\mathcal{X}) \mapsto P_\alpha(G, d, \mathcal{X}), \quad (2)$$

with a particular focus on μ being a counting measure and complexity measurement.

4. Hypothesis

Although previous studies reported that GAN replication is not observed during the normal training procedures in common datasets, the replication itself should be theoretically possible. Intuitively, when the complexity of a

model exceeds data complexity, the model could overfit to the training dataset.

Let us consider an extreme case of a single image training set. Since there is only one image, as long as the the model is complex enough to capture the within-image complexity, it can fit to the training sample. Since there is only one training image, all the latent codes sampled during the training process are set to map to the same image, achieving the replication defined in Equation (1). When the training dataset becomes larger (populated with non-trivial training samples), the empirical distribution of the dataset grows more complex, making overfitting to the dataset more difficult. However, at this stage, the dataset is not small enough for overfitting, but also has not enough samples to reconstruct the underlying image manifold. One can make analogy to the double maximum frequency threshold for lossless reconstruction of a continuous signal in Nyquist-Shannon Theorem [21]. Once the number of training samples is greater than this effective threshold, it becomes possible to reconstruct the image manifold, given a correct model choice, improving the image quality and reducing replication.

Based on the above analysis, we hypothesize that with dataset size increases, the replication probability $P_\alpha(G, d, \mathcal{X})$ decreases. The quality of generated images depicts a U-shape trend with respect to dataset size. The quality is first high when the GAN producing faithful replication of training data. Then both replication and image quality decreases as dataset size increases. Further increasing the dataset size will increase the image quality (photo-realism) while the GAN replication remains low.

5. Definition of Replication

The choice of image distance metric $d(\cdot, \cdot)$ is important for studying GAN replication as it directly defines the meaning of replication. In this study, we use the Euclidean distance in image space as our metric.

This choice might seem counter-intuitive since numerous previous studies [19, 22] explicitly state to avoid using simple Euclidean distance in nearest neighbour (NN) since it can be easily fooled by simple unperceivable color shift, small translation of the image and even dead pixels. Instead, studies have suggested using semantic spaces (like from InceptionV3 deep feature pretrained on ImageNet) for more effective evaluation and memorization check [27, 20, 16].

However, the aforementioned weaknesses of using the Euclidean distance in pixel space is actually a strength in our study. In previous studies, the aim is to show that the proposed GAN does *not* memorize. If the Euclidean metric is used, then as long as the GAN does not produce exact copies of training samples, one can claim the GAN does not merely memorize, which might be too liberal in most applications. However, in our study, the purpose is to show that GANs *do* memorize. Thus using Euclidean distance means replications have to be exact copies (up to a noise level α), which will also be treated as memorization in all the other semantic based metrics.

6. Dataset Size and Complexity

We use two set functions μ_1, μ_2 to characterize training set in our study. The first set function μ_1 is to measure dataset complexity. We use intrinsic dimensionality (ID) for this purpose which can be understood as an estimate of the degree-of-freedom of data manifold in the high dimensional pixel space. We use a maximum likelihood estimator of intrinsic dimensionality [14], which can be defined as,

$$\mu_1(\mathcal{X}) = \frac{1}{|\mathcal{X}|(k_1 - k_2 + 1)} \sum_{k=k_1}^{k_2} \sum_{\mathbf{X} \in \mathcal{X}} \hat{m}_k(\mathbf{X}) \quad (3)$$

$$\hat{m}_k(\mathbf{X}) = \left[\frac{1}{k-1} \sum_{j=1}^{k-1} \log \frac{T_k(\mathbf{X})}{T_j(\mathbf{X})} \right]^{-1} \quad (4)$$

where $T_k(\mathbf{X})$ denotes the Euclidean distance from $\mathbf{X}$ to its k-th NN in $\mathcal{X}$. k_1 and k_2 denote the minimum and maximum k used in k-NN, which affects the locality during ID estimation.

The reason that we define ID in pixel space rather than in any semantic embedding space learnt by a neural network (as in [8]) is to match our choice of $d(\cdot, \cdot)$.

The second set function is the counting measure for dataset size,

$$\mu_2(\mathcal{X}) = |\mathcal{X}|. \quad (5)$$

7. Selection of GAN Architectures

Since we are only interested in studying the relationship between GAN replication and dataset size/complexity, any well-established GAN architecture is reasonable for our purpose. We choose to use StyleGAN2 [13] and BigGAN [2] for their state-of-the-art performance.

We follow the original training procedure of BigGAN and StyleGAN2. For BigGAN implementation, we use BigGANdeep architecture with adversarial hinge loss, which corresponds to the implementation with highest performance in [2].

For StyleGAN2, we use *config-f* in the original paper, which also provides the highest reported performance. The StyleGAN2 *config-f* uses large networks with regularized adversarial logistic non-saturation loss.

8. Dataset Setup

To estimate the f shown in Equation (2), we need to vary the training sets $\mathcal{X}$. To study the effect of dataset size, we

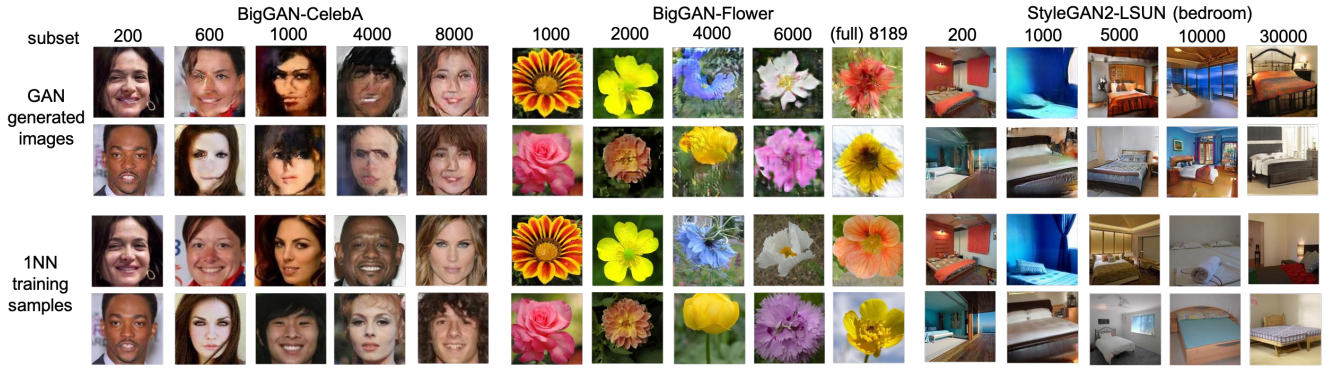

Figure 2: Qualitative results of replication experiments for BigGAN-CelebA, BigGAN-Flower and StyleGAN2-LSUN (bedroom) combination. All images are randomly generated without cherry-picking. Results for all the other experiments are provided in Supplemental Material. For a given GAN and a dataset, at each subset level, a GAN is trained and examined for its replication. This results show that when the dataset size is small, GANs can generate almost exact replication of training data. The replication is gradually alleviated when the dataset size increases.

Datasets	Subset size
CelebA	200*, 600*, 1,000, 4,000, 8,000
LSUN-bedroom	200, 1,000, 50,00, 10,000, 30,000†
Flower	1,000, 2,000*, 4,000, 6,000*, 8,189

Table 1: Subset levels used in our experiments for different datasets. *Subset level only used in BigGAN experiments. †Subset level only used in StyleGAN2 experiments.

starts with a small subset of the training data and gradually increase the dataset size to observe the change in GANs replication $P_\alpha(G, d, \mathcal{X})$. To study the effect of the dataset complexity, we uses multiple datasets for different objects type, combining with various sizes for each set, we build a set of training datasets with different complexities and study their relationship with GAN replication.

The specific dataset used in our study are CelebA [15], LSUN-bedroom [29], Oxford Flower 17 [18]. This collection of training dataset provides a wide range of variety on the dataset complexity, from simple images like faces to complex scenes like bedrooms.

From each dataset, we create multiple levels of subsets with different number of samples randomly selected. The specific number of samples depends on the GAN replication trend on the dataset. The sizes of subset levels used in our study are shown in Table 1.

9. Experimental Setup

To examine the relationship between the dataset size/complexity and GAN replication, we trained BigGAN and StyleGAN2 on each of the subset levels defined in Section 8.

The images from Flower, CelebA and LSUN-bedroom datasets are first center-cropped and scaled to 128×128 resolution. The RGB channles of training images are z-score normalized.

To examine the GAN replication, 1,024 samples are first generated for each trained generator. Given a generated sample, we find its NN in the corresponding training set with Euclidean distance in the original pixel space. Then the percentage of generated samples whose NN distance $< \alpha$ is used as the estimate of $P_\alpha(G, d, \mathcal{X})$. We report results for $\alpha = 8{,}000$ in the main paper, which is the loosest threshold we found consistent with human perception of replication. Additionally, results for $\alpha = 9{,}000$, $10{,}000$ are provided in Supplemental Material to show the effect of different α values on $P_\alpha(G, d, \mathcal{X})$.

For each subset, we calculate the maximum likelihood estimate of its ID [14][1] with $k_1 = 10$ and $k_2 = 20$. Before calculating ID, we down-scale the image to 32×32 as it shorten the runtime with no significant difference comparing to the 128×128 version. The empirical evidence supporting this down-scale operation is also provided in the Supplemental Material.

Both ID and GAN replication calculation require the use of a NN algorithm. We use Faiss [10][2] for its highly efficient implementation of exact NN.

9.1. Measuring perceptual quality

To measure the perceive quality of the generated images, we run a behavioral experiment with human subjects on Amazon Mechanical Turk. One may wonder the necessity of running a behavioral experiment while the evaluation metric such as Fréchet Inception Distance (FID) is avail-

[1] we use a python implementation https://gist.github.com/mehdidc/8a0bb21a31c43b0cbbdd31d75929b5e4/.

[2] https://github.com/facebookresearch/faiss

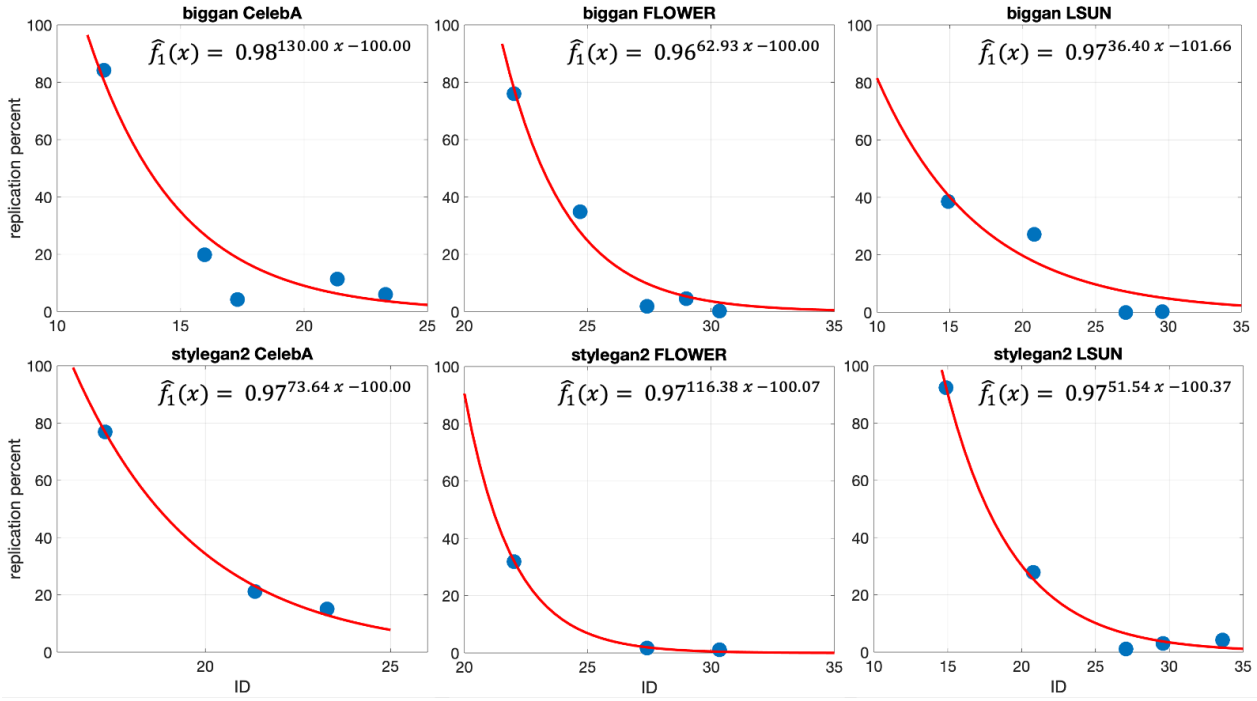

Figure 3: Scatter plots and curve fitting for dataset ID vs GAN replication percentage at each subset level for BigGAN and StyleGAN2 trained on CelebA, Flower and LSUN-bedroom. Regardless of GAN architecture or dataset, the results show a common exponential decay trend. Among the three model parameters to be estimated, the exponential decay factor a and predictor translation c are both shared across GAN architecture and datasets.

able and well accepted. The reason for not using FID is that it does not necessarily reflects the perceived image quality from human subjects. The supporting evidence for this decision can also be found in the Supplemental Material.

We first randomly sampled 100 images per subset level per GAN per dataset. Each image is rated by 9 subjects for its image quality. A 5-point (Excellent-Good-Fair-Poor-Terrible) Likert scale is used. The rating criteria is described in the Supplemental Material.

10. Result and Analysis

Our hypothesis described in Section 4, is that when dataset size increases, the dataset complexity increases (before it converges to the underlying complexity of the ground-truth distribution), and the replication percentage decreases. Additionally, we hypothesize that the quality of generated samples first decreases then increases with increasing number of training samples.

10.1. Dataset complexity vs. GAN replication

Figure 3 shows the relationship between dataset complexity and GAN replication percentages for StyleGAN2 and BigGAN trained on CelebA, LSUN-bedroom and Flower datasets. Each point in the figure represents an experiment on a subset level of the corresponding dataset.

GAN	Datasets	$R^2_{f_1}$	R^2_g	$R^2_{f_2}$
BigGAN	Flower	0.9739	0.9995	0.9709
StyleGAN2	Flower	0.9994	0.9995	0.9996
BigGAN	CelebA	0.9388	0.9999	0.8949
StyleGAN2	CelebA	0.9965	0.9999	0.9956
BigGAN	LSUN	0.8612	1.0000	0.8577
StyleGAN2	LSUN	0.9930	1.0000	0.9922

Table 2: Goodness-of-fit measurement R^2 for the ID-replication function f_1, intermediate ID-size function g and size-replication function f_2 when fitted to our data at all subset levels. The high R^2 shows that all the three formulation of f_1, g and f_2 are highly effective on modeling the relationship between dataset size/complexity and GAN replication.

Although the initial replication percentages are different across datasets, the relationship between dataset complexity and GAN replication percentages follows a common trend of exponential decay across all datasets and GAN architectures.

To quantitatively characterize this trend, we fit an expo-

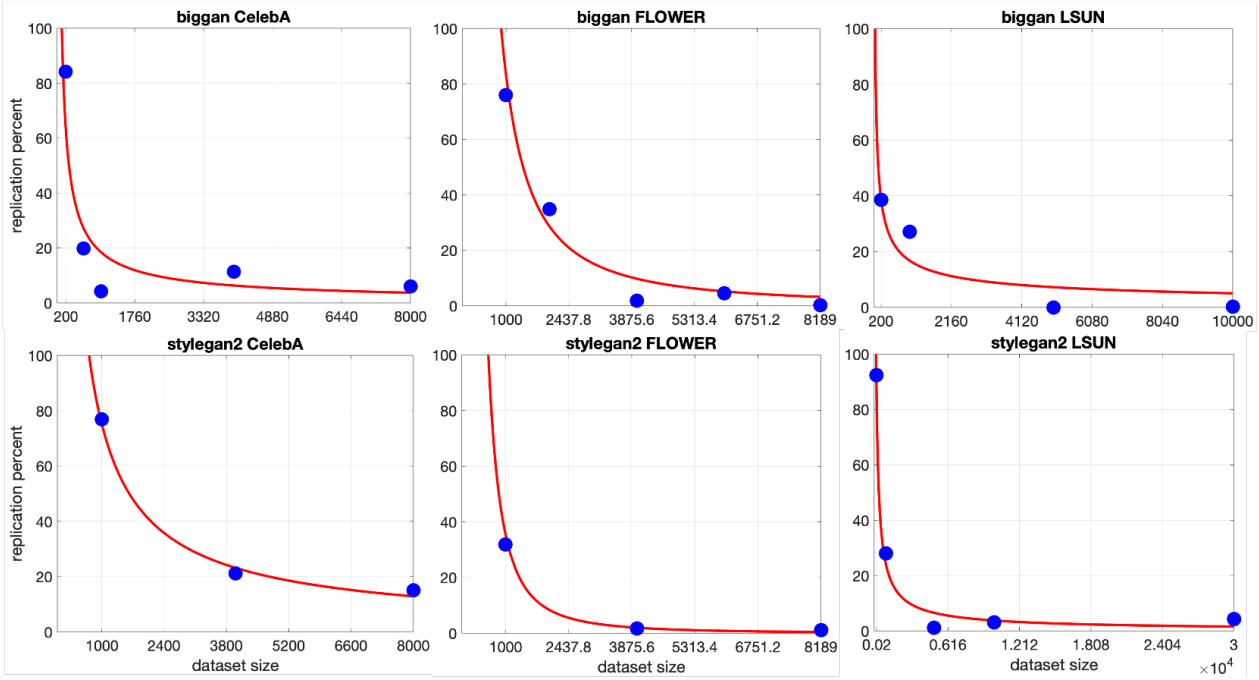

Figure 4: Scatter plots and curve fitting for dataset size vs GAN replication percentage at each subset level. Regardless of GAN architectures or datasets, the results also show a common exponential decay trend similar to the previous ID v.s. replication relationship.

nential function,

$$P_\alpha(G, d, \mathcal{X}) = f_1(\mu_1(\mathcal{X})) = a^{b\mu_1(\mathcal{X})-c} \tag{6}$$

to each of the dataset experiments (with all the subset levels). Parameters a,b and c serves as decay base factor, scaling and translation on the predictor $\mu_1(\mathcal{X})$, respectively. Note that although one may fit the data equally well with $f(x) = a^{x-c}$ with less parameters to estimate, the 3 variable formulation we used here delineates the effect of a and b, which will be an important factor to enable one-shot prediction on GAN replication as shown in Section 11.

Figure 3 also shows that the estimated parameters $\hat{a}$, $\hat{b}$ and $\hat{c}$ of Equation (6) fitted to each of GAN-dataset combinations. Despite with different datasets and GAN architectures, the complexity-replication curves share similar exponential decay factor a and predictor translation c (which can also be understood as response scaling). The former falls in to the range of 0.96 to 0.98 and latter at almost exactly 100.0. With only one parameter b to be determined, we can then predict the full curve when there is only one subset level available for the dataset, which will be shown in Section 11.

We also provide measurement of goodness-of-fit ($R^2_{f_1}$) of our f_1 formulation when fitted to all the subset levels in Table 2, which are greater 0.9 in almost all the experiments, supporting the effectiveness of our f_1 formulation.

10.2. Dataset size vs. GAN replication

Figure 4 shows the relationship between dataset size and GAN replication percentages for StyleGAN2 and BigGAN trained on CelebA, LSUN and Flower datasets. Each point in the figure represents an experiment on a subset level of the corresponding dataset.

Since we already have f_1 defined in Section 10.1 that maps the dataset complexity measurement $\mu_1(\mathcal{X})$ to replication percentage $P_\alpha(G, d, \mathcal{X})$, only a change of variable is needed to define the function f_2 which maps from dataset size $\mu_2(\mathcal{X})$ to GAN replication percentages. We define an intermediate function $g : \mu_1(\mathcal{X}) \mapsto \mu_2(\mathcal{X})$ modeling the relationship between the dataset size and complexity. For g, we formulate it a natural exponential function,

$$\mu_2(\mathcal{X}) = g(\mu_1(\mathcal{X})) = \alpha e^{\beta\mu_1(\mathcal{X})}. \tag{7}$$

As shown in Table 2 R^2_g column, when fitted to the full subset levels of all three datasets, the model yields R^2 close to 1 in all experiments, showing its effectiveness.

Combining Equation 6 and inverse of g from Equation 7, we have f_2 as,

$$\begin{aligned} f_2(\mu_2(\mathcal{X})) &= f_1(g^{-1}(\mu_2(\mathcal{X}))) \\ &= a^{(b/\beta)\ln(\mu_2(\mathcal{X})/\alpha)-c} \end{aligned} \tag{8}$$

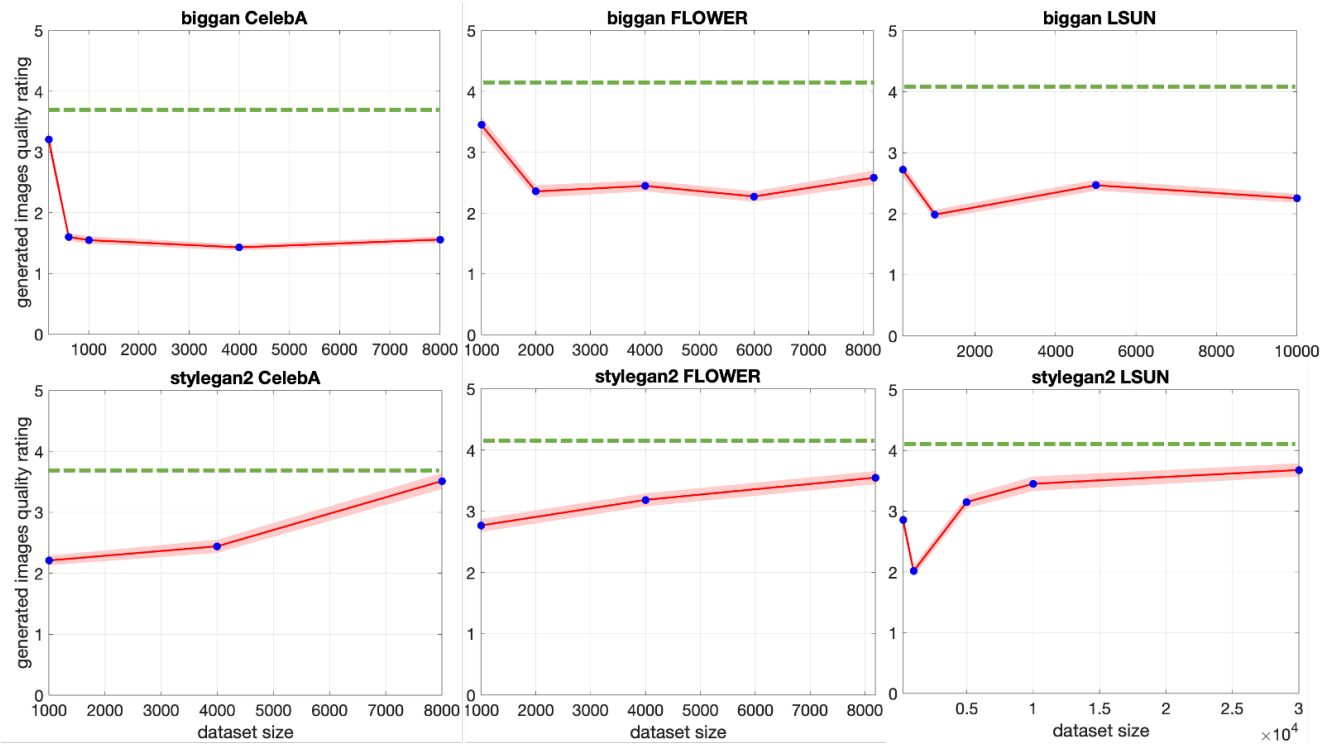

Figure 5: Results of the AMT behavioral experiment testing the relationship between dataset size and perceptual image quality. Average perceptual quality of the image with 95% confidence interval is provided. Green dashed line indicates the average perceptual quality for real images in the dataset. The image quality is high when the dataset size is small and GANs producing exact replication of the training data, except for the StyleGAN2-CelebA and StyleGAN2-Flower experiments.

for each of the dataset experiments (with all the subset levels). The solid red curve in Figure 4 shows the the model when fitted to all the subset levels per GAN-dataset experiment, with the goodness-of-fit provided in Table 2 $R^2_{f_2}$ column. Similar to previous results, our formulation is highly accurate with $R^2 > 0.85$ for all the experiments.

10.3. Dataset size vs. Perceptual Quality

Figure 5 shows the relationship between dataset size and the perceptual quality of the images generated by the GAN trained on the dataset.

Among all combination, the BigGAN-Flower, BigGAN-LSUN and StyleGAN2-LSUN clearly shows the full trend described in our hypothesis which is the perceived image quality is first high when the replication happens, then decrease when the dataset size increases but not large enough, finally increase again when the dataset size becomes larger further.

On the other hand, BigGAN-CelebA, StyleGAN2-CelebA, StyleGAN2-Flower shows partial trend in our hypothesis, with BigGAN-CelebA shows the first half of the trend and StyleGAN2-CelebA, StyleGAN2-Flower showing the second half. This might because even larger subset level is needed for the former and smaller subset level is needed for the latter.

GAN	Datasets	$R^2_{f_1}$		
		full	1-shot	2-shot
BigGAN	Flower	0.9739	0.8228	0.8230
StyleGAN2	Flower	0.9994	1.0000	1.0000
BigGAN	CelebA	0.9388	0.8301	0.8174
StyleGAN2	CelebA	0.9965	0.9958	0.9957
BigGAN	LSUN	0.8612	0.8830	0.8798
StyleGAN2	LSUN	0.9930	0.9925	0.9925

Table 3: Prediction results for complexity-replication function f_1 in one-shot ($R^2_{f_1}$ 1-shot) and two-shot ($R^2_{f_1}$ 2-shot) setups. The result for using full subset levels ($R^2_{f_1}$ full) is provided for reference.

11. One-shot Prediction on GAN Replication

As described in Section 1, a practical purpose of this study is to provide guidance on the dataest size when a new dataset for image synthesis is under construction. This can be achieved by predicting the replication percentage with

GAN	Datasets	MAE_{f_1}(%)		MAE_{f_2}(%)		$\text{MAE}_{f_2^{-1}}$(# of samples)	
		full	1-shot	full	1-shot	full	1-shot
BigGAN	Flower	2.8855	1.8424	6.2709	4.4886	6.0206e2	2.1008e3
StyleGAN2	Flower	0.2144	0.9725	0.6265	1.0143	2.2244e2	1.0546e3
BigGAN	CelebA	5.1180	12.0495	7.4701	12.1064	2.1400e3	2.1982e4
StyleGAN2	CelebA	1.8955	1.2442	2.0183	2.1006	4.5370e2	7.4419e2
BigGAN	LSUN	6.0261	7.3125	6.0358	7.6807	2.4715e6	1.4775e7
StyleGAN2	LSUN	2.4826	3.0008	4.2719	4.6232	3.0231e3	8.6447e2
median		2.6840	2.4215	5.1538	4.5558	1.3710e3	1.5777e3

Table 4: Median Absolute Errors (MAE) of predicting GAN replication from dataset ID (MAE_{f_1}), dataset size (MAE_{f_2}) and predicting dataset size from replication ($\text{MAE}_{f_2^{-1}}$).

dataset size/complexity or vice-versa. Ideally, we wish to predict the curve as early as possible (*i.e.* smallest subset level possible) since the dataset is typically collected by adding more images over time.

11.1. From dataset ID to replication

Section 10.1 shows that the exponential decay factor a and predictor translation c are shared across experiments, leaving only one parameter b to be estimated. Thus it is possible to estimate f_1 for an unseen dataset with as low as one subset level, by using the shared a and c.

To test the ability to predict replication curve for unseen dataset and GAN architectures, we perform an leave-one-out cross-validation (LOOCV). For each GAN-dataset combination shown in Figure 3, we hold out one combination for testing, using the rest to estimate the shared parameter a and c by averaging $\hat{a}$ and $\hat{c}$ across combination. Then, for the held-out combination, we estimate b using only one smallest subset level (one-shot) or first two smallest levels (two-shot). This procedure simulates the practical application when the current dataset under collection is still at its early stage with very small number of samples.

Table 3 shows the goodness-of-fit measurement of predicted GAN replication curve with one-shot and two-shot setup denoted as $R^2_{f_1}$(1-shot) and $R^2_{f_1}$(2-shot) respectively. For comparison, we also include $R^2_{f_1}$(full) for the model fitted with full subsets levels. The table shows that even predicted from only one subset level, the model fits data very well with all $R^2 > 0.8$. Comparing to using full subsets, the one-shot prediction only suffers minor performance drop for BigGAN on Flower and CelebA datasets. The two-shot prediction does not improve over one-shot.

Table 4 MAE_{f_1} shows a more interpretable performance, Mean Absolute Error (MAE) when predicting replication percentage from dataset ID. The results indicate that overall the median error on replication percentage for a given query dataset ID is around 2.42% with only one subset level needed, which does not deviate significantly from the 2.42% median error using full subsets, showing the effectiveness of our one-shot prediction method.

11.2. From dataset size to replication

We perform the same LOOCV one-shot test on predicting replication percentage from dataset size using Equation (8). Table 4 MAE_{f_2} shows the results of this experiment. With the median error around 4.56%, the results show that dataset size is less accurate on predicting replication percentage in our method, which is not surprising due to accumulation of error during the change of variable.

11.3. From replication to dataset size

Since Equation (8) is invertible, we can also predict dataset size for a given replication percentage.

$$f_2^{-1}(x) = \alpha e^{(\beta/b)[\log_a(x)+c]} \tag{9}$$

Table 4 $\text{MAE}_{f_2^{-1}}$ shows the results of this experiment. The mean error is around 1.3K for full subset levels and 1.5K for one-shot prediction, which is relatively small considering the size of modern datasets often range from hundreds of thousands to millions. In this experiment, BigGAN-CelebA and BigGAN-LSUN performs poorly due to large subset level when replication percentage values are close to zero.

12. Discussion and Conclusion

In this study, we show that for a given GAN model and synthesis task, GAN replication percentage decays exponentially w.r.t. dataset size and complexity while the image quality depicts an U-shape trend. We also designed a practical tool to predict the number of training sample necessary in one-shot for a given replication percentage or vice-versa, providing guidance on the choice of dataset size for anyone constructing a novel dataset for image synthesis purpose. This discovery of dataset size and replication relationship also deepen our understanding on the underlying mechanism for GAN replication and overfitting.

References

[1] Sanjeev Arora and Yi Zhang. Do gans actually learn the distribution? an empirical study. *arXiv preprint arXiv:1706.08224*, 2017.

[2] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018.

[3] Ciprian A Corneanu, Sergio Escalera, and Aleix M Martinez. Computing the testing error without a testing set. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2677–2685, 2020.

[4] Ciprian A Corneanu, Meysam Madadi, Sergio Escalera, and Aleix M Martinez. What does it mean to learn in deep networks? and, how does one detect adversarial attacks? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4757–4766, 2019.

[5] Georgios Douzas and Fernando Bacao. Effective data generation for imbalanced learning using conditional generative adversarial networks. *Expert Systems with applications*, 91:464–471, 2018.

[6] Maayan Frid-Adar, Idit Diamant, Eyal Klang, Michal Amitai, Jacob Goldberger, and Hayit Greenspan. Gan-based synthetic medical image augmentation for increased cnn performance in liver lesion classification. *Neurocomputing*, 321:321–331, 2018.

[7] Jessica L Gillotte. Copyright infringement in ai-generated artworks. *UC Davis L. Rev.*, 53:2655, 2019.

[8] Sixue Gong, Vishnu Naresh Boddeti, and Anil K Jain. On the intrinsic dimensionality of image representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3987–3996, 2019.

[9] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[10] Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with gpus. *arXiv preprint arXiv:1702.08734*, 2017.

[11] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017.

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

[13] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8110–8119, 2020.

[14] Elizaveta Levina and Peter J Bickel. Maximum likelihood estimation of intrinsic dimension. In *Advances in neural information processing systems*, pages 777–784, 2005.

[15] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

[16] Mario Lucic, Karol Kurach, Marcin Michalski, Sylvain Gelly, and Olivier Bousquet. Are gans created equal? a large-scale study. In *Advances in neural information processing systems*, pages 700–709, 2018.

[17] Vaishnavh Nagarajan, Colin Raffel, and I Goodfellow. Theoretical insights into memorization in gans.

[18] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *Indian Conference on Computer Vision, Graphics and Image Processing*, Dec 2008.

[19] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

[20] Mehdi SM Sajjadi, Olivier Bachem, Mario Lucic, Olivier Bousquet, and Sylvain Gelly. Assessing generative models via precision and recall. In *Advances in Neural Information Processing Systems*, pages 5228–5237, 2018.

[21] Claude Elwood Shannon. Communication in the presence of noise. *Proceedings of the IRE*, 37(1):10–21, 1949.

[22] Lucas Theis, Aäron van den Oord, and Matthias Bethge. A note on the evaluation of generative models. *arXiv preprint arXiv:1511.01844*, 2015.

[23] Patrick Tinsley, Adam Czajka, and Patrick Flynn. This face does not exist... but it might be yours! identity leakage in generative models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1320–1328, 2021.

[24] Ngoc-Trung Tran, Viet-Hung Tran, Ngoc-Bao Nguyen, Trung-Kien Nguyen, and Ngai-Man Cheung. On data augmentation for gan training. *IEEE Transactions on Image Processing*, 30:1882–1897, 2021.

[25] Fei Wang, Zhanyao Zhang, Chun Liu, Yili Yu, Songling Pang, Neven Duić, Miadreza Shafie-Khah, and João PS Catalão. Generative adversarial networks and convolutional neural networks based weather classification model for day ahead short-term photovoltaic power forecasting. *Energy conversion and management*, 181:443–462, 2019.

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

[27] Qiantong Xu, Gao Huang, Yang Yuan, Chuan Guo, Yu Sun, Felix Wu, and Kilian Weinberger. An empirical study on evaluation metrics of generative adversarial networks. *arXiv preprint arXiv:1806.07755*, 2018.

[28] Yasin Yazici, Chuan-Sheng Foo, Stefan Winkler, Kim-Hui Yap, and Vijay Chandrasekhar. Empirical analysis of overfitting and mode drop in gan training. In *2020 IEEE International Conference on Image Processing (ICIP)*, pages 1651–1655. IEEE, 2020.

[29] Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015.

[30] Bingquan Zhu, Hao Fang, Yanan Sui, and Luming Li. Deepfakes for medical video de-identification: Privacy protection and diagnostic information preservation. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 414–420, 2020.