Commonality in Natural Images Rescues GANs: Pretraining GANs with Generic and Privacy-free Synthetic Data

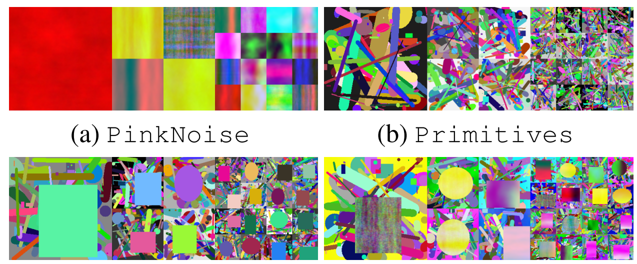
Kyungjune Baek, Hyunjung Shim†
Yonsei University
{bkjbkj12, kateshim}@yonsei.ac.kr

Abstract

Transfer learning for GANs successfully improves generation performance under low-shot regimes. However, existing studies show that the pretrained model using a single benchmark dataset is not generalized to various target datasets. More importantly, the pretrained model can be vulnerable to copyright or privacy risks as membership inference attack advances. To resolve both issues, we propose an effective and unbiased data synthesizer, namely `Primitives-PS`, inspired by the generic characteristics of natural images. Specifically, we utilize 1) the generic statistics on the frequency magnitude spectrum, 2) the elementary shape (i.e., image composition via elementary shapes) for representing the structure information, and 3) the existence of saliency as prior. Since our synthesizer only considers the generic properties of natural images, the single model pretrained on our dataset can be consistently transferred to various target datasets, and even outperforms the previous methods pretrained with the natural images in terms of Fréchet inception distance. Extensive analysis, ablation study, and evaluations demonstrate that each component of our data synthesizer is effective, and provide insights on the desirable nature of the pretrained model for the transferability of GANs.

1. Introduction

Generative adversarial networks (GANs) [13] are a powerful generative model that can synthesize complex data by learning the implicit density distribution with adversarial training. Thanks to the impressive generation quality, particularly in image generation tasks [4, 23, 30], GANs have been widely used in various downstream tasks in computer vision, such as data augmentation [9], super-resolution [25,54], image translation [1,10], and image synthesis with primitive representation [27,37]. Despite the remarkable quality, GANs require at least several thousand, mostly several hundred thousand images for training. This requirement for data collection is often infeasible in practical applications (e.g., many pictures of a treasure, endangered species, or the medical images of rare disease).



(a) `PinkNoise` (b) `Primitives` (c) `Primitives-S` (d) `Primitives-PS`

Figure 1. Visualization of our synthetic datasets. We visualize four variants of our synthetic datasets and `Primitives-PS` is finally chosen for the best performance. Example images are resized in three different scales.

The idea of transfer learning has been recently introduced to GAN training [31, 49] for resolving the real-world generation problem. Following the common practice, the general framework of GAN transfer learning 1) pretrains GANs on a publicly available large-scale source dataset (e.g., FFHQ and ImageNet) and then 2) finetunes GANs with a relatively small target dataset. As a result, developing GANs with transfer learning clearly improves the generation quality and diversity over the models trained from scratch only with the target dataset.

Unfortunately, the effectiveness of transfer learning for GANs highly depends on how similar the source dataset is to the target dataset. According to TransferGAN [49], transfer learning can achieve the best performance when the source shares common characteristics with the target. For example, when LFW [21] is the target dataset, the best performance is achieved with the source dataset of CelebA [28] as both are face datasets. For Flower [33] or Kitchens [53], utilizing CelebA as the source dataset does not significantly improve the performance. Thus, it is required to search the best source dataset for a given target dataset by measuring the similarity between two datasets (e.g., FID score). Be-

†Hyunjung Shim is a corresponding author.

cause exploring the best source dataset and then acquiring its pretrained model is ad-hoc, the search result does not guarantee the best pair for transfer learning [49]. Moreover, none of the existing source datasets can sufficiently fit the target dataset in real-world applications.

Other than the performance issue, we argue that the current pretrained models can be vulnerable to copyright (see the supplementary 7 for potential copyright issues of large-scale datasets) and privacy issues [58]. Even for public benchmark datasets, employing them for commercial purposes is not always permitted. For examples, ImageNet-1K having 1M images, the copyright issue might not be feasible to handle. When targeting the commercial use of a dataset, the developer should negotiate with the author of each sample. For this reason, one might compose her own dataset via web crawling, but filtering out the copyrighted samples is practically difficult. Besides, unresolved copyright and privacy issues might cause legal issues [42].

Recent studies [8, 15, 18] also show that the deep generative models are vulnerable to membership inference attacks, implying that privacy issues still remains beyond the copyright issues. An adversary can reconstruct a face even without additional prior information [55]. That is, we can reveal individual training samples by attacking the trained model. As the network capacity of GANs increases rapidly to improve performance, the risk of memorization also grows quickly. Memorization effects make GANs more vulnerable to membership inference attacks [7]. Since we consider transfer learning, someone might argue that the membership inference on the source (e.g., pretraining) dataset should not be a critical issue. However, Zou et al. [58] reported that the membership inference of the source dataset could be conducted even after the transfer learning (see the supplementary 7 for empirical evidence).

In this work, we dive into tackling the two undiscovered but critical issues of transfer learning for GANs: 1) the lack of generalization for the pretrained model and 2) the copyright or privacy issue of the pretraining dataset. To this end, we devise a synthetic data generation strategy for acquiring pretrained GANs. Since our pretrained model is newly computed with a synthetic dataset, it is inherently free from copyright and privacy issues. Besides, the learned features of existing pretrained models encode the inductive bias of a training dataset, exhibiting lower transferability [52]. Learned from this lesson, we ensure that our synthetic data should be unbiased to any datasets and free from expert knowledge or specific domain prior.

Towards this goal, we adopt the generic property of the natural images in the frequency spectrum and structure. We develop our data generation strategy, namely `Primitives-PS`, inspired by the analysis and observations on natural images from previous studies [29, 36, 44]. Our design philosophy is built upon three aspects: 1) considering the power spectrum distribution of the natural images as in Figure 1(a), 2) reflecting the structural property of the natural images as illustrated in Figure 1(b), and 3) utilizing the existence of saliency in images (Figure 1(c) shows the synthetic data generated by applying both 2) and 3).) Finally, we combine all three aspects and develop our final data synthesizer `Primitives-PS`, as visualized in Figure 1(d). We pretrain GANs using the synthetic dataset generated by our data synthesizer. Then, the effectiveness of the proposed method is evaluated by repurposing the pretrained model to various low-shot datasets.

Extensive evaluations and analysis confirm that this single pretrained network 1) can be effectively transferred to various low-shot datasets and 2) improve the generation performance and the convergence time. Interestingly, the model pretrained with our dataset outperforms the model pretrained with the natural images when transferred to several datasets. Our empirical study shows that the bias from a specific dataset for pretraining GANs is harmful to the generalization performance of transfer learning. Finally, our analysis of learned filters provides insight into what makes the pretrained model transferable. The code is available at https://github.com/FriedRonaldo/Primitives-PS.

2. Related work

2.1. Utilizing synthetic datasets

The samples and labels of synthetic datasets can be generated automatically and unlimitedly by a pre-defined process. Since generating synthetic data can bypass the cumbersome data crawling and pruning for data collection, previous works have utilized synthetic datasets for training the model and then achieved performance improvement on real datasets [19, 20, 39–41, 45, 51]. Domain randomization [45] used various illuminations, color, noise, and texture to reduce the performance gap between the simulated and real samples. By doing so, a model trained with a synthetic dataset helps improve the performance on the real dataset. Fourier domain adaptation [51] proposed swapping the low-frequency components of the synthetic and real samples to reduce the domain gap in the texture.

Although the previous methods improved the performance of the model on the real dataset, generating such synthetic datasets requires expertise in domain knowledge or a specific software (e.g., GTA-5 game engine [38]). To handle the issue, Kataoka et al. [24] utilized the iterated function system to generate fractals and used the fractals as a pretraining dataset for classification. As a concurrent work, Baradad et al. [3] observe that the unsupervised representation learning [16] trains the model using patches, and these patches are visually similar to the noise patches (from the noise generation model) or the patches drawn from GANs. Based on the observation, they generate synthetic datasets and conduct self-supervised learning for an image classification task. However, none of the existing studies have investigated synthetic data generation for training GANs.

2.2. Transfer learning in GANs

GANs involve a unique architecture and a training strategy; consisting of a discriminator and generator trained via adversarial competition. Therefore, the GAN transfer learning method should be developed by considering the unique characteristics of GANs [31, 34, 35, 48, 49, 56]. TransferGAN [49] trains GANs with a small number of samples by transferring the weights trained on a relatively large dataset. TransferGAN also shows that the performance of the transferred model depends on the relationship between the source and target datasets. Noguchi and Harada [34] proposed to update only the statistics of the batch normalization layer for transferring GANs. This strategy prevents GANs from overfitting so that the model can generate diverse images even with a small number of samples. FreezeD [31] fixes several layers of the discriminator and then finetunes the remaining layers. FreezeD improved the generation performance of transferring from the FFHQ pretrained model to various animals. Despite the improvement in GAN transfer learning, the model still requires a large-scale pretraining dataset. Consequently, they commonly suffer from copyright issues, and their performance is sensitive to the relationship between the source and target dataset. In contrast, our goal is to tackle both issues simultaneously by introducing an effective data synthesizer.

2.3. Low-shot learning in GANs

For high-quality image generation, GANs require a large-scale dataset, and such a requirement can limit the practical use of GANs. To reduce the number of samples for training, several recent studies have introduced data augmentation for training the discriminator [22, 47, 57]. Then, the generator can produce images with a small number of samples without reflecting an unwanted transformation such as cutout [11] in the results (i.e., augmentation leakage [22]). Recently, ReMix [6] utilizes interpolation in the style space to reduce the required images to train an image-to-image translation model. In this work, we tackle low-shot generation using GANs via transfer learning; GANs are trained with a small number of samples by transferring a pretrained network into a low-shot dataset.

3. Towards an effective data synthesizer

In this work, our primary goal is to develop an unbiased and effective data synthesizer. The synthetic dataset secured by our synthesizer is then used to pretrain GANs, which facilitates low-shot data generation. To accomplish unbiased data generation, we only consider the generic properties of natural images because the inductive bias in a pretraining dataset is harmful to transfer learning of GANs. In the following, we introduce three design philosophies of our data synthesizer inspired by the common characteristics of natural images: 1) learning the power spectrum of natural images, 2) exploiting the shape primitives from natural images, and 3) adopting the existence of saliency in images.

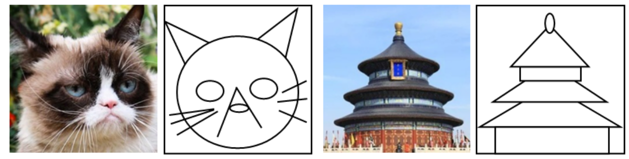


Figure 2. Potentials of primitive shapes for representing things. We only use a line, ellipse, and rectangle to express a cat and a temple. These examples motivate us to develop `Primitives`, which generates the data by a simple composition of the shapes.

3.1. Learning the power spectrum of natural images

Several previous works reported the magnitude of natural images in the frequency domain [5, 12, 46] roughly obeys $w_m = \frac{1}{|f_x|^a + |f_y|^a}$ where a is a constant, well approximated to one. Inspired by this finding, we generate synthetic images by randomly drawing a from the uniform distribution of $\mathcal{U}(0.5, 3.5)$, as also suggested in [3]. Specifically, random white noise is sampled, and then its magnitude signal after applying the Fast Fourier Transform (FFT) is weighted by w_m. By applying the inverse FFT to the weighted signal, we can easily compute the synthetic image. We repeat this for RGB color channels and finally produce synthetic images. Originally, the image with $a = 1$ was named a pink noise. We call this method of generating images with $a \sim \mathcal{U}(0.5, 3.5)$ as `PinkNoise`. Since we only utilize the generic properties of natural images, no inductive bias toward any specific dataset influences `PinkNoise`. As shown in Figure 1(a), `PinkNoise` produces interesting patterns with vertical, horizontal orientation, or color blobs.

3.2. Shape primitives inspired by natural images

"Everything in nature is formed upon the sphere, the cone, and the cylinder. One must learn to paint these simple figures, and then one can do all that he may wish."

Paul Cézanne

Considering the importance of phase in images (e.g., determining the unique appearance of the image [36]), `PinkNoise` alone is insufficient to represent the rich characteristics of natural images; `PinkNoise` is random noise on a phase spectrum. To have a meaningful signal even in its phase, we can consider 1) modeling the phase of natural images independently or 2) developing the different generation strategies to model the magnitude and phase simultane-

Table 1. SSIM between the magnitude spectrum of the frequency domain of the synthetic and target dataset. The higher score means the more similar pair. We observed that the tendency is the same with L1 or L2 distance.

Source \ Target	Obama	Grumpy cat	Bridge	Panda	FFHQ	Mean
`PinkNoise`	0.8368	0.8148	0.7676	0.8328	0.8553	0.8215
`Primitives`	0.9309	0.9366	0.9198	0.9200	0.9635	0.9342
`Primitives-S`	0.9421	0.9463	**0.9308**	0.9334	0.9756	0.9456
`Primitives-PS`	**0.9432**	**0.9476**	0.9307	**0.9352**	**0.9767**	**0.9467**

ously. Unlike the magnitude spectrum, we seldom find regularity in the phase of images; thus, it is difficult to derive the generic property of the phase spectrum. Besides, separately modeling the phase and magnitude may not produce meaningful images, preserving the proper structures [44]. For this reason, we focus on finding structural regularity in natural images because it can affect both magnitude and phase. Specifically, we are inspired by the observation that natural images can be represented by the composition of the elementary shapes [29]. The common practice in artistic drawings also utilizes elementary shapes as the basis for representing things (inspired by Paul Cézanne).

Figure 2 demonstrates the abstraction examples of various images using elementary shapes, such as ellipses, lines, and rectangles. We find the potential of abstraction via elementary shapes to encode the structural information of natural images and to remove the bias to a specific dataset. We then devise the data synthesizer to produce images consisting of various elementary shapes. The outputs of this synthesis procedure are akin to those of the dead leaves model [14, 26]. The dead leaves model is an early generative model, which closely mimics natural images by conducting tessellation, where their sizes and positions are determined by sampling from the Poisson process. Unlike the dead leaves model, we do not fill all the regions and use different distributions for sampling because the resultant images are quite sensitive to the hyperparameter of the Poisson process. For position, we use the uniform distribution. To prevent the large shapes in the later stage from completely overwriting those in the early stage, we gradually decrease the maximum shape size over multiple stages; drawing the small objects toward the end. In addition, it is conversely proportional to the number of currently injected shapes. We name this generation strategy `Primitives`, and Figure 1(b) visualizes the representative examples. By distributing the shapes in the image space, we observe that `Primitives` produces images that have a similar magnitude to those of natural images (See Table 1 and the supplementary 10 for the supporting experiments).

3.3. Combining saliency as prior

In addition to the natural images, we investigate the benchmark datasets and find that they commonly have saliency, target objects of interest to determine the class. These salient objects are usually located nearly in the middle of the image. For example, the animal face on the cat and panda dataset can be the saliency. To reflect the nature of benchmark datasets, we insert a large shape after applying `Primitives` and name it as `Primitives-S` (`Primitives` with Saliency).

(a) (b)

Figure 3. Comparison between (a) `Primitives` and (b) `Primitives-PS` on Obama dataset. The model pretrained with `Primitives` generates multiple faces in a single image.

By utilizing the three design factors, we develop four variants of our data synthesizer. They are 1) `PinkNoise` adopting the nature of magnitude spectrum of natural images only as shown in Figure 1(a), 2) `Primitives` generating various elementary (monotone) shapes randomly as illustrated in Figure 1(b), and 3) `Primitives-S` adding a salient object into `Primitives` in Figure 1(c).

Finally, we apply a `PinkNoise` pattern onto the salient object and the background of `Primitives-S`, which is called (4) `Primitives-PS` (`Primitives` with Patterned Saliency) as shown in Figure 1(d). Since the size of the salient object is considerable, having a salient monotone object may induce an unwanted texture bias. Focusing on the visual effects, inserting the monotone object can be similar to the regional dropout [2, 43] in the weakly-supervised object localization (WSOL) task. When training a network with the regional dropout, previous WSOL methods suggest filling the dropped region with mean statistics or with other regions from the same image to prevent distribution bias. Motivated by the practice in WSOL, we apply `PinkNoise` to the salient object.

The effectiveness of the proposed synthetic datasets is evaluated by transferring GANs in Section 4. We first pretrain GANs using the randomly generated images via our `Primitives-PS`, and then finetune the pretrained model on low-shot datasets. While finetuning, all competitors and our pretrained model utilize DiffAug (translation, cutout, and color jittering). For the pretraining results and the details, please refer to the supplementary 9.

Table 2. The FID score of transferring to low-shot datasets from the proposed pretraining datasets. The lower is the better. Bold and underlined text indicates the best and second best performance among the pretraining datasets. It will be the same convention throughout the paper.

Source \ Target	Obama	Grumpy cat	Bridge	Panda
Scratch + DiffAug	48.98	27.51	57.72	15.82
`PinkNoise`	50.32	29.47	73.82	15.65
`Primitives`	<u>43.20</u>	27.97	59.89	12.78
`Primitives-S`	43.29	<u>26.57</u>	<u>57.24</u>	**11.95**
`Primitives-PS`	**41.62**	**26.01**	**54.02**	<u>12.23</u>

4. Experiments

We first demonstrate the effectiveness of four variants of our data synthesizer. Then, we choose the best strategy among the four variants and use it for pretraining GANs. Our pretrained model is compared with other pretrained models using a natural benchmark dataset in the transfer learning scenario. We also provide an ablation study on the number of particles in each synthetic image and a policy to determine the size of each particle in the supplementary 1.

Datasets. For the comparison between our synthesizers, we adopt four datasets, including Obama, Grumpy cat, Panda, and Bridge of sighs (Bridge) [57]. To compare with transfer learning methods, we also use Wuzhen, Temple of heaven (Temple), and Medici fountain (Fountain). Each dataset has 100 images. In addition, we create a dataset, namely Buildings, by merging a subset of four datasets; Bridge of sighs, Wuzhen, Temple of heaven, and Medici fountain. Buildings is used to evaluate the performance under highly diverse conditions. For comprehensive evaluations, we also use CIFAR-10/100 datasets when training with BigGAN.

Evaluation protocols. StyleGAN2 architecture [23] with DiffAug [57] is applied when evaluating all models in the low-shot generation task. The baseline is the model trained from scratch with DiffAug. The strong competitors are TransferGAN [49] and FreezeD [31], where both methods suggest finetuning strategies. To reproduce the competitors, we pretrain StyleGAN2 on FFHQ– the face dataset and then finetune the pretrained model using TransferGAN with DiffAug and FreezeD with DiffAug, respectively. Since the baseline can outperform the competitors upon the target datasets, we report the baseline performances for comparison. Besides, we stress that all competitors, baseline and `Primitives-PS` use DiffAug. Specifically, we follow the configuration of DiffAug for `Primitives-PS` and the baseline (from scratch with DiffAug). Otherwise, we use the configuration of TransferGAN and FreezeD as described in [57] for the best performance.

We also apply our synthetic dataset to pretrain BigGAN [4] and repurpose the model to CIFAR-10/100 datasets for evaluating our synthesizer in the conditional generation task. Since `Primitives-PS` does not have labels, we randomly assigned the labels during pretraining. We developed the pretrained model independently for CIFAR-10 and 100 as they have different architectures due to different numbers of classes. For evaluating the conditional generation task, we compare three models; 1) the model naïvly trained from scratch, 2) the model trained with DiffAug only (DiffAug), and 3) our model pretrained with `Primitives-PS` and then finetuned with DiffAug. We use 10%, 20%, and 100% samples of CIFAR for evaluation and check the effectiveness of our strategy under the data-scarce scenario. As an evaluation metric, we use Fréchet inception distance (FID) [17] and report the FID score of the best model during training as suggested by DiffAug [57]. We also provide KMMD [50] for the better quantitative evaluation, please refer to supplementary 11.

Figure 4. Morphing upon the transfer learning iterations of the `Primitives-PS` pretrained model. We generate the images by using the same latent vector. The center lilac circles are gradually changed into salient regions.

4.1. Effects of different data synthesizers

We developed four variants of data synthesizer: `PinkNoise`, `Primitives`, `Primitives-S`, and `Primitives-PS`. We evaluate their effectiveness in the low-shot generation scenario– pretraining with the synthetic dataset and then finetuning on target datasets with DiffAug. Table 2 summarizes the FID scores of four data synthesizers and the baseline under four different low-shot datasets.

In general, `PinkNoise` fails to improve the FID score (worse than the baseline), but converges fast (See the supplementary 2). Unlike `PinkNoise`, `Primitives` clearly improves the generation performance in Obama and Panda, large margins from the baseline. However, it is not effective on Grumpy cat and Bridge. Compared to `Primitives`, `Primitives-S` further improves the FID scores, demonstrating the effectiveness of saliency prior. Finally, `Primitives-PS` clearly improves the low-shot generation performance on all datasets by about 15% on average over the baseline. We provide the qualitative evaluation in the supplementary 3. From these results, we observe that 1) a naïve synthesizer (`PinkNoise`) is even worse than simply using the low-shot dataset, and 2) the combi-

(a) From scratch (b) TransferGAN (c) FreezeD (d) `Primitives-PS`

Figure 5. Qualitative evaluation on Obama, Grumpy cat, Temple, and Wuzhen. For more results, please refer to the supplementary 5.

Table 3. The FID score of transferred models to low-shot datasets. We use FFHQ pretrained weight for TransferGAN and FreezeD. For all models, we apply DiffAug. Bold and underlined text indicates the best and second best performance among the pretraining datasets.

Source \ Target	Obama	Grumpy cat	Bridge	Panda	Temple	Wuzhen	Fountain	Buildings
Scratch + DiffAug [57]	48.98	<u>27.51</u>	<u>57.72</u>	15.82	46.69	146.81	<u>44.46</u>	93.71
TransferGAN [49]	<u>36.50</u>	30.60	60.29	14.53	<u>40.58</u>	95.83	46.61	81.63
FreezeD [31]	**35.90**	29.41	59.47	<u>13.39</u>	42.09	<u>93.54</u>	45.70	<u>80.48</u>
`Primitives-PS`	41.62	**26.01**	**54.02**	**12.23**	**40.42**	**88.14**	**43.06**	**78.74**

nation of our three design factors (`Primitives-PS`) remarkably improves the baseline, supporting the effectiveness and importance of each factor.

To analyze how closely our data synthesizers mimic the real datasets, we focus on measuring the similarity between our synthetic dataset (source) and the actual low-shot dataset (target). Instead of pixel distance, we compare the average structural similarity (SSIM) between two datasets in the frequency domain. Since the phase periodically varies in $[-\pi, \pi]$, the SSIM of the phase spectrum is not reliable for comparison. Therefore, we only report the SSIM using the magnitude spectrum in Table 1. We confirm that similar trends are consistently observed in L1 or L2 distance. The value of the SSIM is not an exact indicator for explaining the FID scores. Nevertheless, it helps understand the gains; the low-shot generation performance improves as our data synthesizer models the target dataset more similarly. In Table 2, `Primitives-S` and `Primitives-PS` were ranked top-2, except for Obama. The two strategies in Table 1 also show that their magnitude spectrum is the most similar to target datasets. This interesting trend supports that our design factors are effective choices to mimic the statistics of real images.

We also visualize how our synthetic data gradually fit the target data by showing the generation results at different training stages. For that, `Primitives` and `Primitives-PS` are selected to construct the pretrained model, and then they are transferred to Obama. By comparing `Primitives` and `Primitives-PS`, we observe the effect of the saliency prior. Figure 3 shows that the salient shape in `Primitives-PS` forms the main object as the training evolves. Meanwhile, `Primitives` includes multiple shapes, meaning all can be candidates for the main object. Consequently, the results often contain multiple faces in the middle of training (e.g., the top-left, the top-right, and the middle in Figure 3(a)). On the other hand, `Primitives-PS` focuses on generating a single face and eventually exhibits improved quality. We further visualize the gradual changes in outputs of `Primitives-PS` pretrained model in Figure 4. For the full animation, please refer to the supplementary material (GIF files).

Considering all, we confirm that `Primitives-PS` is the best data synthesizer, and thus it is chosen as our final model for comparative evaluations with competitors.

4.2. Comparisons with the state-of-the-arts

We pretrain a model using `Primitives-PS` and compare it with state-of-the-art models pretrained with natural images in a transfer learning task to low-shot datasets.

Table 3 reports the quantitative results and Figure 5 shows the qualitative comparison. As expected, TransferGAN [49] and FreezeD [31] show outstanding performance on the Obama dataset because they are pretrained with FFHQ, meaning the source dataset is a superset of the target. Except for the Obama dataset, our pretrained model with `Primitives-PS` outperforms all competitors. Unless the inductive bias in the source dataset is advantageous to the target (e.g., Obama), FreezeD does not consistently outperform the baseline (from scratch with DiffAug). In fact, the performances of existing methods highly vary upon target datasets. Contrarily, our pretrained model with `Primitives-PS` consistently outperforms

Table 4. The average consine similarity between the filters in the same layer. The lower value indicates the more diverse filters.

Pretraining DB	Discriminator	Generator
`Primitives-PS`	**0.00820**	**0.00828**
FFHQ	0.01348	0.01434

Table 5. The FID of BigGAN, with DiffAug, and with DiffAug initialized by `Primitives-PS` (`PS`) pretrained model on CIFAR. '*' indicates the best FID before augmentation leakage [22]. Please refer to the supplementary 8 for the details.

	CIFAR-10			CIFAR-100		
	10%	20%	100%	10%	20%	100%
BigGAN	44.14	20.80	9.45	66.21	34.78	13.45
+ DiffAug	29.78*	14.04	**8.55**	41.70*	21.14	11.51
+ Pretrained (`PS`)	**21.33**	**12.79**	8.79	**32.57**	**20.58**	**11.29**

the competitors in each dataset, except Obama. This implies that our pretrained model has strong transferability. Since `Primitives-PS` does not use any inductive bias for modeling human faces, the performance drawback on Obama can be acceptable.

We emphasize that our achievement in generation quality is impressive and meaningful in two aspects: 1) `Primitives-PS` uses no real but all synthetic images, which possesses all the attractive nature in application scenarios and 2) our results show the great potential of a single pretrained model for GAN transfer learning.

Diverse filters matter for transferring GANs. From the superior performances of our pretrained model, we conjecture that our achievement was possible by the unbiased nature of our dataset; the pretrained model with FFHQ (FreezeD) has an inductive bias as the face dataset. A previous study analyzing the transferability of CNN [52] also pointed out that the performance of the target dataset degrades when the filters are highly specialized to the source dataset. To analyze the transferability empirically, we measure the similarity between the filters of each layer of the pretrained model. We regard that highly diverse (less similar to each other) filters can indicate that the model is less biased towards a particular domain. That means that the highly transferable model tends to have low filter similarity on average. Specifically, given a weight matrix of each layer, its shape is $[O, I, H, W]$, where O filters have $I \times H \times W$ tensors. Then, we measure the cosine similarity among all possible permutations of O filters and report the mean value of the average similarity of all layers in Table 4. For all the layers, please refer to the supplementary 6.

In summary, `Primitives-PS` shows the more diverse filter set in 21 out of 26 layers than the FFHQ pretrained model. According to [52], the higher layer (close to the output) tends to specialize in the trained dataset. The same observation holds in our discriminator. The similarity in the last layer of the FFHQ pretrained model is approximately four times higher than `Primitives-PS`. This explains that the FFHQ pretrained model specialized in human faces,

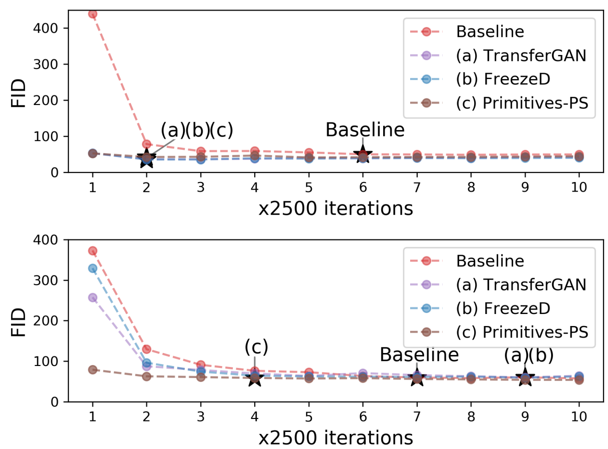
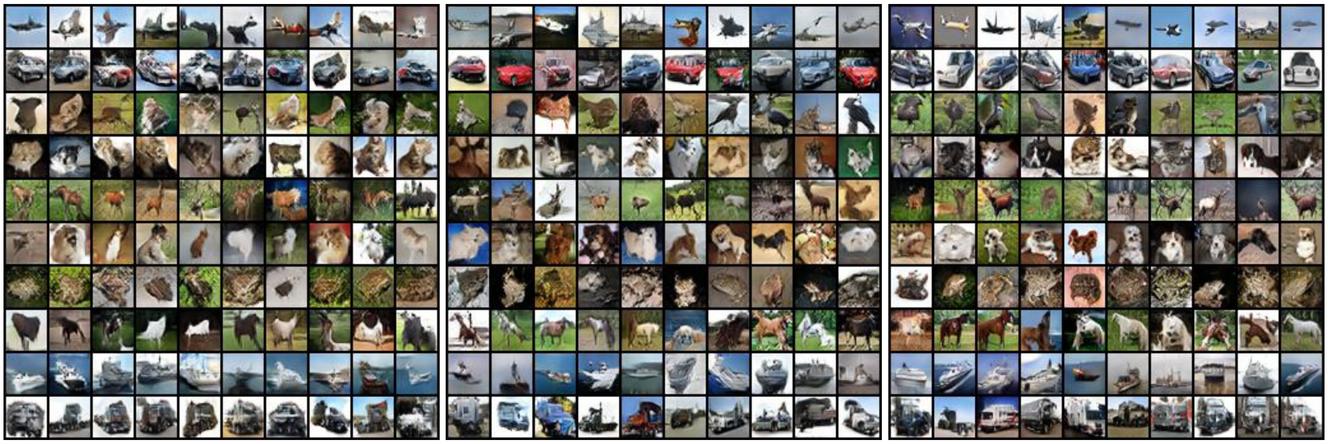


Figure 6. FID per training iterations. The star marker (★) indicates the point where the model reaches 95% of the best FID score of the from scratch model with DiffAug (baseline). Our `Primitives-PS` pretrained model is comparable to the competitors on Obama dataset (upper) and converges faster than the others on Bridge of sighs dataset (lower).

thus transferring well to Obama but not to others.

Training convergence speed. We investigate the convergence speed of transfer learning by examining FID upon training iterations. Figure 6 describes the evolution of the FID scores during the training. To save space, we provide two different datasets; Obama and Bridge. Results for the complete set are in the supplementary 4. For Obama, all pretrained models converge faster than the baseline (from scratch with DiffAug). Meanwhile, only our model converges faster than the baseline for Bridge. Compared to the baseline, the model pretrained with `Primitives-PS` reaches 95% of the best baseline performance within the first 30% of iterations. Interestingly, other pretrained models cannot reach 95% of the best baseline performance earlier than the baseline. This shows that our model effectively reduces the required iterations for convergence, and the overhead for pretraining can be sufficiently deducted.

Toward a conditional generation task using CIFAR. We conduct conditional generation via transfer learning on CIFAR-10 and 100 as summarized in Table 5. Figure 7 shows the qualitative evaluation result on CIFAR-10 with 10% of samples; our `Primitives-PS` produces the general shape and its structural components better than the baseline and DiffAug. Compared to BigGAN trained from scratch, BigGAN trained from scratch with DiffAug significantly improves the FID score, and the gain is pronounced as the number of training samples decreases. However, we observe that DiffAug suffers from augmentation leakage [22] when the samples are scarce (i.e., the generated samples contain the cutout). Our pretrained model with `Primitives-PS` shows remarkable performances under the data-hungry scenario, better than DiffAug.



(a) From scratch (b) + DiffAug (c) + `Primitives-PS` pretraining

Figure 7. Qualitative evaluation on CIFAR-10 dataset with 10% of samples. Each row contains samples in the same class.

However, when the samples are sufficient (100%), pretraining does not always provide gains over DiffAug. This tendency appears in various downstream tasks. Newell et. al. [32] reported that the self-supervised pretraining for semi-supervised classification is not advantageous when the amount of data-label pairs are sufficient. TransferGAN [49] showed that the gain via transfer learning decreases when the amount of samples is sufficient. In the same vein, the advantage of our pretraining with `Primitives-PS` decreases as the number of samples increases.

For the extreme low-shot scenario, we also evaluated the model trained with 1% of the dataset. Only for this evaluation, we compare three models; 1) the model naïvly trained from scratch, 2) the model trained with DiffAug only (DiffAug), and 3) our model pretrained with `Primitives-PS` and then finetuned without DiffAug. The FID score of the baseline, DiffAug, and ours are 112.13, 101.91, and 78.48, respectively. Although DiffAug improved FID, we observe that DiffAug suffers from the augmentation leakage issue. Therefore, the improvement in FID and its generation results are not meaningful. In contrast, our pretrained model can significantly improve the generation performance without any issue. For more details and results for CIFAR, please refer to the supplementary 8.

5. Discussion and conclusion

Societal impact. Since we propose the synthetic dataset for pretraining, the proposed method can consume more power at the pretraining stage (generating the synthetic data and then pretraining the model). However, it converges much faster for finetuning on target datasets, and the same model can be repeatedly used for all targets. In this regard, our method is eventually the more efficient choice in terms of power consumption. In the point of the ethical view, especially considering the bias issues (e.g., racial or gender bias) in the current benchmark datasets, using our method is much more safe, fair, economical, and practical. Besides, pretraining with our synthetic dataset guarantees the robustness of membership inference attack towards the source dataset because reconstructing our data is meaningless. Since our method is copyright-free, it helps small commercial groups to develop their machine-learning model.

Limitation. Our `Primitives-PS` is devised based on the observations from natural images. Hence, it is possible that more effective observations can further improve the data generation quality. In future work, we plan to develop a metric to quantify the transferability of the model and then derive the data generation process by optimizing the transferability. Formulating such a metric will be challenging but constructive for predicting the behavior of the pretrained model and practically useful in various applications.

Conclusion. Existing studies for GAN transfer learning utilize a model trained with natural images and thereby suffer from 1) biased pretrained model that can be harmful to the resultant performance and 2) copyright or privacy issues with both the model and dataset. To overcome these limitations, we introduce a new image synthesizer, namely `Primitives-PS`, inspired by the three generic properties of natural images: 1) following the power spectrum of natural images, 2) abstracting the image via the composition of primitive shapes (e.g., line, circle, and rectangle), and 3) having saliency in the image. Experimental comparisons and analysis show that our strategy effectively improves both the generation quality and the convergence speed. We further investigate the diversity of learned filters and report that they are meaningful evidence for discovering the transferability of the pretrained model.

Acknowledgements. We thank Jongwuk Lee and CVML members for the valuable feedback. This research was supported by the NRF Korea funded by the MSIT (2022R1A2C3011154, 2020R1A4A1016619), the IITP grant funded by the MSIT (2020-0-01361, YONSEI UNIVERSITY), and the Korea Medical Device Development Fund grant (202011D06).

References

[1] Kyungjune Baek, Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Hyunjung Shim. Rethinking the truly unsupervised image-to-image translation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14154–14163, 2021. 1

[2] Kyungjune Baek, Minhyun Lee, and Hyunjung Shim. Psynet: Self-supervised approach to object localization using point symmetric transformation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 10451–10459, 2020. 4

[3] Manel Baradad, Jonas Wulff, Tongzhou Wang, Phillip Isola, and Antonio Torralba. Learning to see by looking at noise. In *Advances in Neural Information Processing Systems*, 2021. 2, 3

[4] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natural image synthesis. In *International Conference on Learning Representations*, 2019. 1, 5

[5] Geoffrey J Burton and Ian R Moorhead. Color and spatial structure in natural scenes. *Applied optics*, 26(1):157–170, 1987. 3

[6] Jie Cao, Luanxuan Hou, Ming-Hsuan Yang, Ran He, and Zhenan Sun. Remix: Towards image-to-image translation with limited data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15018–15027, 2021. 3

[7] Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos, and Dawn Song. The secret sharer: Evaluating and testing unintended memorization in neural networks. In *28th {USENIX} Security Symposium ({USENIX} Security 19)*, pages 267–284, 2019. 2

[8] Dingfan Chen, Ning Yu, Yang Zhang, and Mario Fritz. Gan-leaks: A taxonomy of membership inference attacks against generative models. In *Proceedings of the 2020 ACM SIGSAC conference on computer and communications security*, pages 343–362, 2020. 2

[9] Junsuk Choe, Song Park, Kyungmin Kim, Joo Hyun Park, Dongseob Kim, and Hyunjung Shim. Face generation for low-shot learning using generative adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 1940–1948, 2017. 1

[10] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8188–8197, 2020. 1

[11] Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017. 3

[12] David J Field. Relations between the statistics of natural images and the response properties of cortical cells. *Josa a*, 4(12):2379–2394, 1987. 3

[13] 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. 1

[14] Robert M Haralick, Stanley R Sternberg, and Xinhua Zhuang. Image analysis using mathematical morphology. *IEEE transactions on pattern analysis and machine intelligence*, (4):532–550, 1987. 4

[15] Jamie Hayes, Luca Melis, George Danezis, and Emiliano De Cristofaro. Logan: Membership inference attacks against generative models. In *Proceedings on Privacy Enhancing Technologies (PoPETs)*, volume 2019, pages 133–152. De Gruyter, 2019. 2

[16] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020. 2

[17] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 5

[18] Benjamin Hilprecht, Martin Härterich, and Daniel Bernau. Monte carlo and reconstruction membership inference attacks against generative models. *Proc. Priv. Enhancing Technol.*, 2019(4):232–249, 2019. 2

[19] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei Efros, and Trevor Darrell. Cycada: Cycle-consistent adversarial domain adaptation. In *International conference on machine learning*, pages 1989–1998. PMLR, 2018. 2

[20] Yunzhong Hou and Liang Zheng. Visualizing adapted knowledge in domain transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13824–13833, 2021. 2

[21] Gary B Huang and Erik Learned-Miller. Labeled faces in the wild: Updates and new reporting procedures. *Dept. Comput. Sci., Univ. Massachusetts Amherst, Amherst, MA, USA, Tech. Rep*, 14(003), 2014. 1

[22] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 12104–12114. Curran Associates, Inc., 2020. 3, 7

[23] 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. 1, 5

[24] Hirokatsu Kataoka, Kazushige Okayasu, Asato Matsumoto, Eisuke Yamagata, Ryosuke Yamada, Nakamasa Inoue, Akio Nakamura, and Yutaka Satoh. Pre-training without natural images. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*, November 2020. 2

[25] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on*

computer vision and pattern recognition, pages 4681–4690, 2017. 1

[26] Ann B Lee, David Mumford, and Jinggang Huang. Occlusion models for natural images: A statistical study of a scale-invariant dead leaves model. *International Journal of Computer Vision*, 41(1):35–59, 2001. 4

[27] Jianan Li, Tingfa Xu, Jianming Zhang, Aaron Hertzmann, and Jimei Yang. LayoutGAN: Generating graphic layouts with wireframe discriminator. In *International Conference on Learning Representations*, 2019. 1

[28] 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. 1

[29] Ravish Mehra, Qingnan Zhou, Jeremy Long, Alla Sheffer, Amy Gooch, and Niloy J Mitra. Abstraction of man-made shapes. In *ACM SIGGRAPH Asia 2009 papers*, pages 1–10. 2009. 2, 4

[30] Lars Mescheder, Andreas Geiger, and Sebastian Nowozin. Which training methods for gans do actually converge? In *International conference on machine learning*, pages 3481–3490. PMLR, 2018. 1

[31] Sangwoo Mo, Minsu Cho, and Jinwoo Shin. Freeze the discriminator: a simple baseline for fine-tuning gans. In *CVPR AI for Content Creation Workshop*, 2020. 1, 3, 5, 6

[32] Alejandro Newell and Jia Deng. How useful is self-supervised pretraining for visual tasks? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7345–7354, 2020. 8

[33] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing*, pages 722–729. IEEE, 2008. 1

[34] Atsuhiro Noguchi and Tatsuya Harada. Image generation from small datasets via batch statistics adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2750–2758, 2019. 3

[35] Utkarsh Ojha, Yijun Li, Jingwan Lu, Alexei A Efros, Yong Jae Lee, Eli Shechtman, and Richard Zhang. Few-shot image generation via cross-domain correspondence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10743–10752, 2021. 3

[36] Alan V Oppenheim and Jae S Lim. The importance of phase in signals. *Proceedings of the IEEE*, 69(5):529–541, 1981. 2, 3

[37] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2337–2346, 2019. 1

[38] Stephan R. Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *European Conference on Computer Vision (ECCV)*, volume 9906 of *LNCS*, pages 102–118. Springer International Publishing, 2016. 2

[39] Subhankar Roy, Evgeny Krivosheev, Zhun Zhong, Nicu Sebe, and Elisa Ricci. Curriculum graph co-teaching for multi-target domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5351–5360, 2021. 2

[40] Kuniaki Saito, Kohei Watanabe, Yoshitaka Ushiku, and Tatsuya Harada. Maximum classifier discrepancy for unsupervised domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3723–3732, 2018. 2

[41] Rui Shu, Hung Bui, Hirokazu Narui, and Stefano Ermon. A DIRT-t approach to unsupervised domain adaptation. In *International Conference on Learning Representations*, 2018. 2

[42] Natasha Singer and Mike Isaac. Facebook to pay $550 million to settle facial recognition suit. *The New York Times*, 2019. 2

[43] Krishna Kumar Singh and Yong Jae Lee. Hide-and-seek: Forcing a network to be meticulous for weakly-supervised object and action localization. In *International Conference on Computer Vision (ICCV)*, 2017. 4

[44] Y Tadmor and DJ Tolhurst. Both the phase and the amplitude spectrum may determine the appearance of natural images. *Vision research*, 33(1):141–145, 1993. 2, 4

[45] Josh Tobin, Rachel Fong, Alex Ray, Jonas Schneider, Wojciech Zaremba, and Pieter Abbeel. Domain randomization for transferring deep neural networks from simulation to the real world. In *2017 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, pages 23–30. IEEE, 2017. 2

[46] DJ Tolhurst, Y_ Tadmor, and Tang Chao. Amplitude spectra of natural images. *Ophthalmic and Physiological Optics*, 12(2):229–232, 1992. 3

[47] 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. 3

[48] Yaxing Wang, Abel Gonzalez-Garcia, David Berga, Luis Herranz, Fahad Shahbaz Khan, and Joost van de Weijer. Minegan: effective knowledge transfer from gans to target domains with few images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9332–9341, 2020. 3

[49] Yaxing Wang, Chenshen Wu, Luis Herranz, Joost van de Weijer, Abel Gonzalez-Garcia, and Bogdan Raducanu. Transferring gans: generating images from limited data. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 218–234, 2018. 1, 2, 3, 5, 6, 8

[50] 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. 5

[51] Yanchao Yang and Stefano Soatto. Fda: Fourier domain adaptation for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4085–4095, 2020. 2

[52] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? *Advances in Neural Information Processing Systems*, 27:3320–3328, 2014. 2, 7

[53] Fisher Yu, Yinda Zhang, Shuran Song, Ari Seff, 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. 1

[54] Wenlong Zhang, Yihao Liu, Chao Dong, and Yu Qiao. Ranksrgan: Generative adversarial networks with ranker for image super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3096–3105, 2019. 1

[55] Yuheng Zhang, Ruoxi Jia, Hengzhi Pei, Wenxiao Wang, Bo Li, and Dawn Song. The secret revealer: Generative model-inversion attacks against deep neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 253–261, 2020. 2

[56] Miaoyun Zhao, Yulai Cong, and Lawrence Carin. On leveraging pretrained gans for generation with limited data. In *International Conference on Machine Learning*, pages 11340–11351. PMLR, 2020. 3

[57] Shengyu Zhao, Zhijian Liu, Ji Lin, Jun-Yan Zhu, and Song Han. Differentiable augmentation for data-efficient gan training. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2020. 3, 5, 6

[58] Yang Zou, Zhikun Zhang, Michael Backes, and Yang Zhang. Privacy analysis of deep learning in the wild: Membership inference attacks against transfer learning. *arXiv preprint arXiv:2009.04872*, 2020. 2