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

Contents

1. Ablation Study

When developing `Primitives-PS`, we introduce two hyperparameters; 1) the total number of shapes and 2) the policy to determine the size of each component. For determining the size, we consider three policies; **Fix**, **Rand** and **Decay**. **Fix** indicates that all particles have the same size. To examine the effect of various scale, we set this size as $H \cdot [1/10, 1/5, 1/2]$, where H is the image resolution. **Rand** randomly samples the size from the uniform distribution. Both policies can induce the occlusion of the previously injected shapes by the later shape. **Decay** can bypass the occlusion issue effectively. **Decay** arbitrarily samples the size from the uniform distribution, where the maximum size is limited to $(H \cdot 1/5 \cdot (N-n)/N)$, and N and n are the total number of shapes and the number of previously injected particles. In this way, we can ensure that the shapes inserted in the early stage are still visible in the final data. The upperside of Table 1 summarizes the FID score for each policy on four datasets. The differences in FID among **Fix** policies are trivial in that their ratios are not highly correlated with their ranks. Also, we observe that the shapes at the final stage overwrite the previous shapes. Then, the overall appearance with **Fix** are similar to `PinkNoise` with a salient object. We investigate the synthesizer that combines `PinkNoise` with `PS` by injecting a saliency and then applying `PinkNoise` on it. Interestingly, we observe that it shows the similar FID scores to **Fix**. For **Rand**, it improves the FID score on Obama and bridge, however, the overall performance is much worse than **Decay**. Therefore, we choose a **Decay** policy as default for choosing the size.

Besides, the total number of shapes is important because it affects the transferability and the time complexity of the synthesizer. The lower-side of Table 1 demonstrates the performance trends upon the total number of shapes. A zero particle case implies that only one background and one salient object, thus equivalent to `PinkNoise` + `PS`. As the number of shapes (N) grows upon roughly 100, the performance tends to improve. However, over $N = 100$, we do not observe the consistent gain. From the ablation study, we decide $N = 100$ in each image to enjoy the reasonable

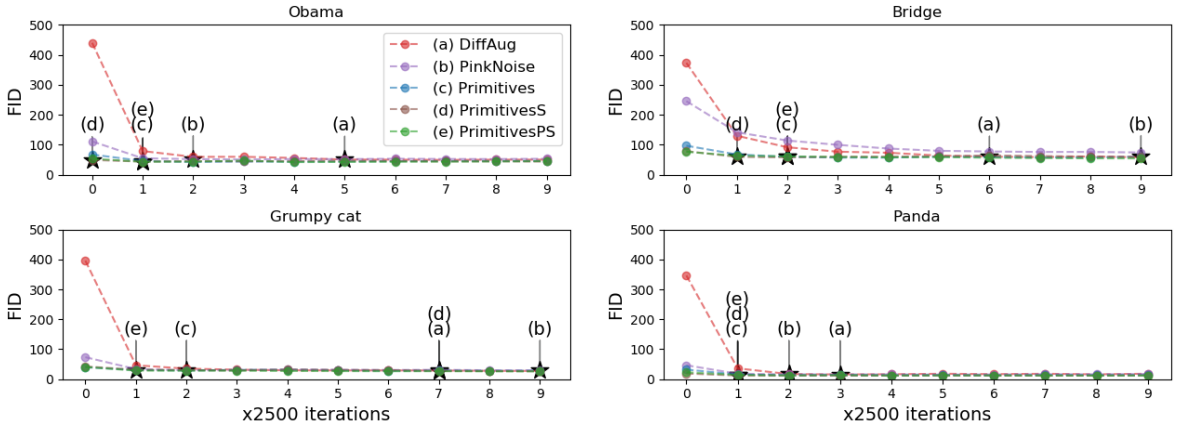


Figure 1. FID per training iterations. The star marker (★) indicates the point where the model reaches 95% of the best FID score of the from scratch model with DiffAug (baseline). The legend is the same for all graphs.

Table 1. Ablation study on the policy to determine the size of each particle (upper) and the number of particles (lower).

Policy	Obama	Grumpy cat	Bridge	Panda
Fix ($1/10$)	48.30	29.74	63.00	17.69
Fix ($1/5$)	46.41	29.22	64.02	14.97
Fix ($1/2$)	48.05	29.37	64.65	15.14
`PinkNoise + PS`	49.13	29.87	66.00	15.12
Rand	44.85	29.84	60.45	14.67
Decay	**41.62**	**26.01**	**54.02**	**12.23**
# of particles	Obama	Grumpy cat	Bridge	Panda
0	49.13	29.87	66.00	15.12
10	44.10	28.00	63.26	13.35
50	42.49	28.40	59.17	**11.79**
100	**41.62**	**26.01**	54.02	12.23
500	42.45	27.92	**52.27**	12.12

performance gain and to reduce the time complexity.

2. Convergence speed of synthetic datasets

Figure 1 shows the evolution of the FID scores during the training of the models pretrained with synthetic datasets. Even if `PinkNoise` does not improve the generation performance, it can boost the convergence speed. In general, the pretrained models reach 95% of the best FID score of the from scratch model with DiffAug within first 30% iterations. The faster convergence speed informs us the positive potential of the pretraining.

3. Qualitative comparison among our data synthesizers

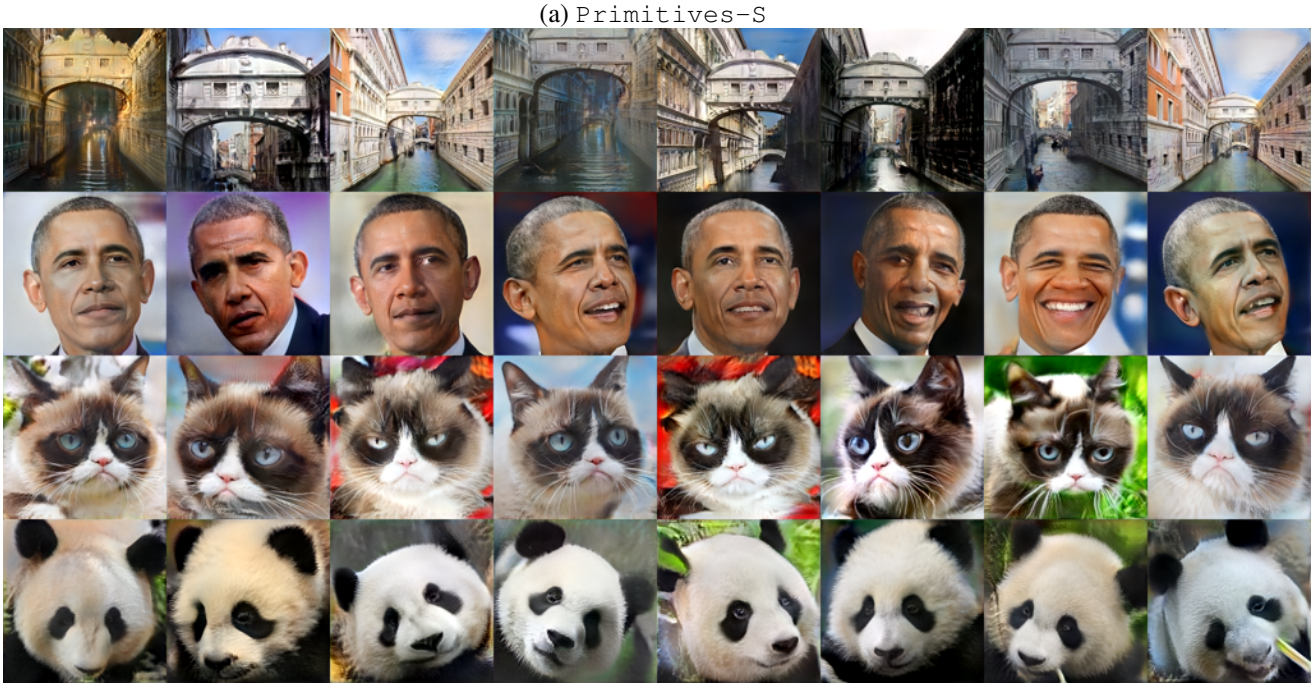
In addition to the quantitative comparison of our data synthesizers, we also qualitatively compare our four variants of the data synthesizer used for quantitative evaluation. From the first to the last row, Bridge of sighs, Obama, Grumpy cat, and Panda. `PinkNoise` generates the images with unstructured samples (e.g. Obama and Grumpy cat) and the outputs of `Primitives` on Panda have lower fidelity (e.g. the last three samples). Compared to `PinkNoise` and `Primitives`, `Primitives-S` and `Primitives-PS` provide plausible samples. Between the last two synthetic datasets, `Primitives-S` sometimes drops the important factor, for example, the eyes of the cat (6-th column). While `Primitives-PS` generates more diverse and plausible samples than the other synthetic datasets.

(a) `PinkNoise`

(b) `Primitives`

Figure 2. Low-shot image generation results of the models transferred from `PinkNoise` and `Primitives`.

(a) `Primitives-S`



(b) `Primitives-PS`

Figure 3. Low-shot image generation results of the models transferred from `Primitives-S` and `Primitives-PS`.

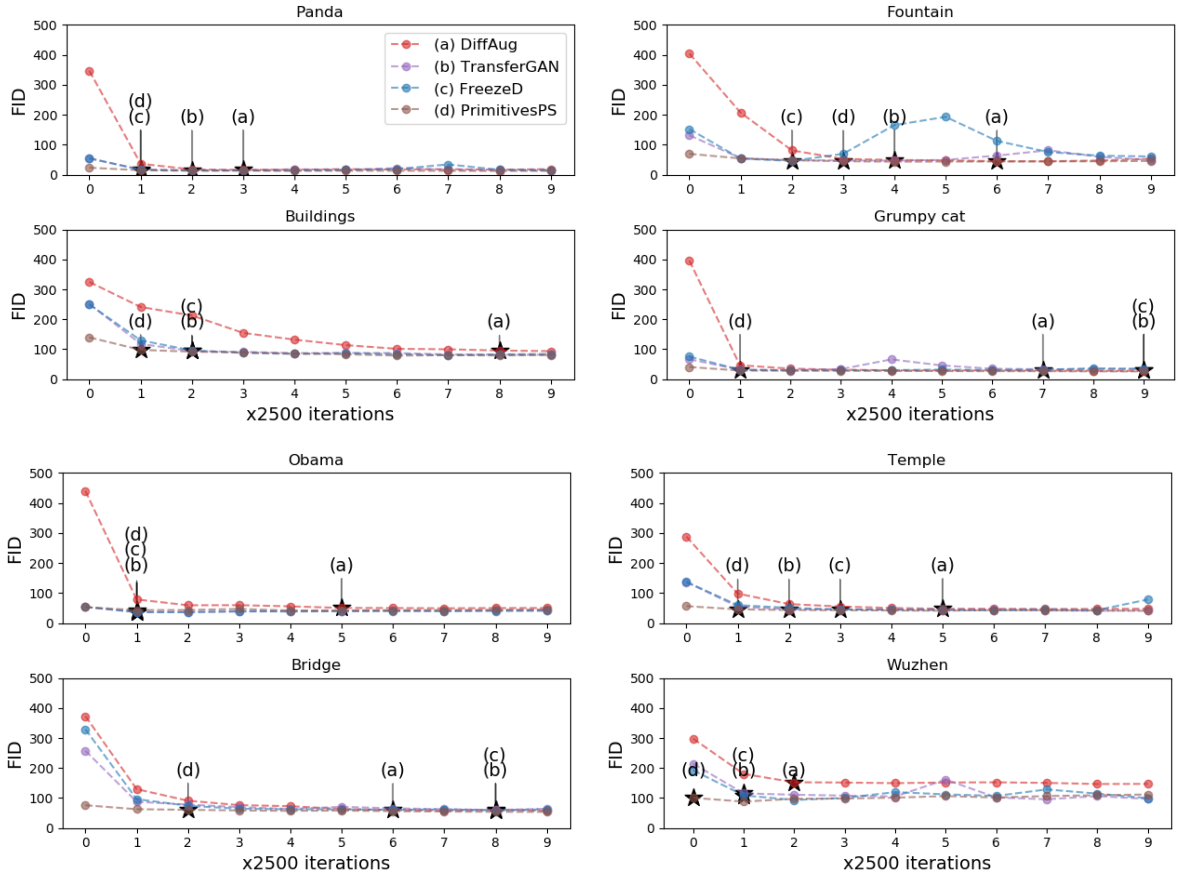


Figure 4. The additional results of Figure 6 in the main text. FID per training iterations. The star marker (★) indicates the point where the model reaches 95% of the best FID score of the from scratch model with DiffAug (baseline). The legend is the same for all graphs.

4. Convergence speed of transfer learning methods

Figure 4 shows the evolution of the FID scores during the training of the transfer learning methods. The model pretrained with our synthetic dataset exhibits comparable or faster convergence than the competitors that are pretrained on FFHQ. Herein, we observe the convergence speed in terms of the number of iterations to reach 95% of the best FID score of the baseline (from scratch model with DiffAug).

5. Qualitative comparisons with competing transfer learning methods

In addition to the quantitative comparison, we also provide the qualitative comparisons on eight datasets that are used for quantitative evaluation in the main text. From the first to the last row, Buildings, Bridge of sighs, Obama, Medici fountain, Grumpy cat, Temple of heaven, Panda, and Wuzhen.

In terms of fidelity of the generated images, our `Primitives-PS` outperforms the competitors. Especially, Grumpy cat images generated by the competitors often do not contain eyes or have only part of the face.

Figure 5. The additional generated samples of Figure 5 in the main text. The images are generated with the model trained from scratch.

Figure 6. The additional generated samples of Figure 5 in the main text. The images are generated with the model pretrained with FFHQ and transferred by using TransferGAN.

Figure 7. The additional generated samples of Figure 5 in the main text. The images are generated with the model pretrained with FFHQ and transferred by using FreezeD.

Figure 8. The additional generated samples of Figure 5 in the main text. The images are generated with the model pretrained with our `Primitives-PS`.

Table 2. The additional results of Table 4 in the main text. The average consine similarity between the filters in the same layer. The lower value indicates the more diverse set of filters.

	Discriminator		Generator	
	Primitives-PS	FFHQ	Primitives-PS	FFHQ
conv0	**0.00660**	0.01245	**0.00315**	0.00685
conv1	0.02104	**0.00932**	**0.00273**	0.00843
conv2	0.01012	**0.00779**	**0.00291**	0.00956
conv3	**0.00839**	0.01216	**0.00348**	0.01080
conv4	**0.00607**	0.00713	**0.00539**	0.01059
conv5	**0.00596**	0.00668	**0.00329**	0.01406
conv6	**0.00507**	0.00563	**0.00363**	0.01199
conv7	**0.00632**	0.00714	**0.00433**	0.01465
conv8	0.00380	**0.00365**	**0.00652**	0.01317
conv9	**0.00521**	0.00703	**0.00933**	0.01626
conv10	0.00503	**0.00420**	**0.01133**	0.01778
conv11	**0.00462**	0.00760	0.01981	**0.01977**
conv12	**0.01844**	0.08438	**0.03176**	0.03250
Mean	**0.00820**	0.01348	**0.00828**	0.01434

6. Similarity between filters in all layers

We calculated the cosine similarity in each layer to measure the diversity of learned filters of pretrained models. FFHQ pretrained model exhibits lower diversity in filters. The average similarity at the last layer of FFHQ pretrained model is approximately four times higher than Primitives-PS. The similar tendency is shown in the first layer of each network – the consine similarity of FFHQ pretrained model is about two times higher than Primitives-PS.

Table 3. Membership inference performance on the source dataset by attacking a transferred classifier as reported in [3].

Dataset	AUC	Accuracy	Precision	Recall
CIFAR100	0.522	0.502	0.478	0.523
Flowers102	0.528	0.496	0.432	0.505
PubFig83	0.495	0.481	0.396	0.524

7. Copyright issue and vulnerability of pretrained model

When we directly finetune a pretrained model for commercial use, the trained weights of the model might be defined as software and have the CC BY-ND (creative commons license without modification) license. In this case, we can not utilize the model with post-training or should pay the license fee for the model as software. If we want to use the images for non-commercial purposes, we should acquire the credit of each image from the original author. For 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. Since this process requires much time and cost to complete, it is likely to be an obstacle to the practical usage of the deep learning system.

Even if we solve the copyright issue via negotiation, the leakage of the training data is another problem. Following the recent work [3], the source dataset for pretraining a model can be exposed by the membership inference attack even after the transfer learning. Table 3 shows the empirical evidence. The target models are first pretrained on Caltech101 and transferred to three datasets. The higher AUC, the higher accuracy of the membership inference on the source dataset. Although the accuracy is lower than the attack on the target dataset, it warns us to consider the membership inference attack towards the source dataset seriously.

Figure 9. Examples of the leakage when using DiffAug. The gray box in some images shows the leakage of cutout operation.

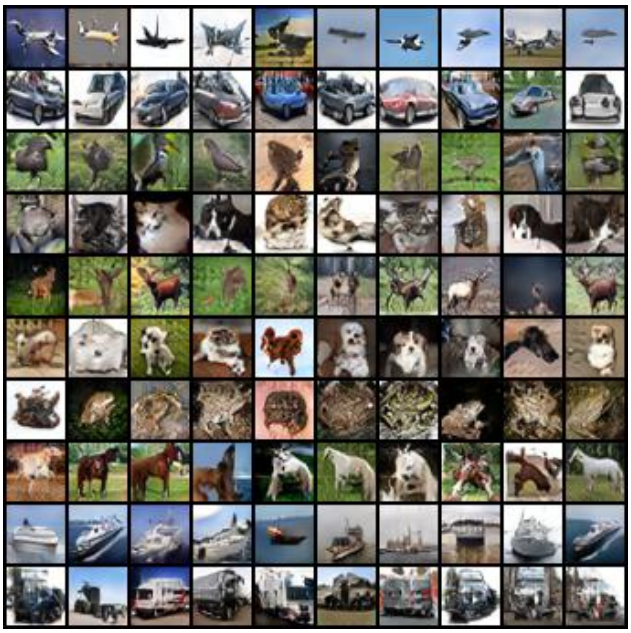


Figure 10. Outputs of the model transferred from our model on CIFAR-10. The model does not suffer from augmentation leakage although we use DiffAug.

8. Experimental results on CIFAR

8.1. Data augmentation leakage

The previous work [1] reported the ill-behavior of the data augmentation in GANs; augmentation leakage. When the leakage incurs, the unwanted data transformation is reflected in the generated results. For example, the generated images contain cutout augmentation so that some of the fakes have unwanted empty box. When we train BigGAN on CIFAR with 10% of samples using DiffAug only, we observe that augmentation leakage. Although the leakage is found, the FID score decreases; FID scores can not reflect the problem of leakage. To penalize this unwanted result, we qualitatively exclude the model with leakage when we find the best model. Figure 9 shows the generated images by the model trained with DiffAug (FID: 22.54). Many of the outputs have the unwanted gray box that is the result of leakage of the cutout operation, and this is why we exclude the corresponding FID score in Table 5 of the main text.

On the contrary, the model pretrained with `Primitives-PS` does not suffer from the leakage even if we use DiffAug (Figure 10). It shows that our pretraining dataset is also effective to prevent augmentation leakage and improves the final generation quality.

9. Pretraining results and details

In this section, we provide the outputs of the generator pretrained with `Primitives-PS`. For pretraining, we train the model during 800K images with batch size = 16, therefore, the total number of iterations is 50K. For finetuing all the models, we train the model during 400K images. The generated (fake) synthetic images are similar to the real synthetic samples as shown in Figure 1 of the main text.

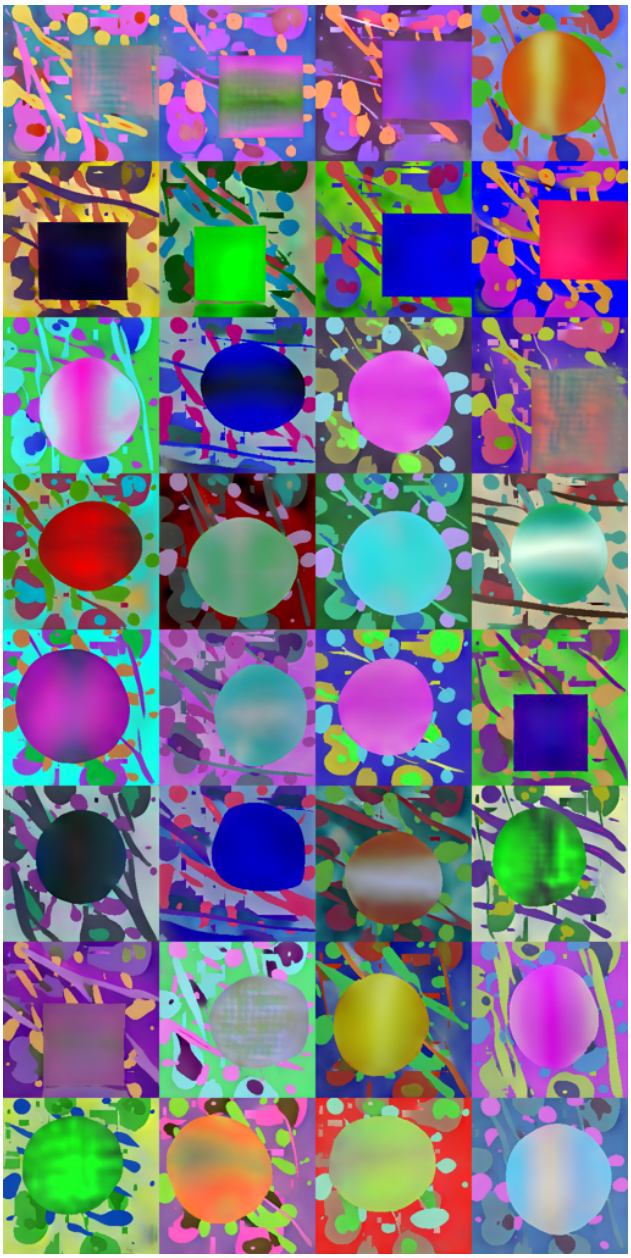
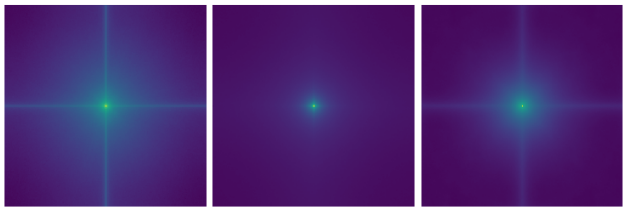


Figure 11. The outputs of the model pretrained with `Primitives-PS`. The generated outputs are similar to the synthetic samples.

10. Frequency domain analysis

We visualize the average magnitude spectrum of all the samples in Bridge of sighs and compare with the average magnitude spectrum of 1000 images generated by `PinkNoise` and 1000 images generated by `Primitives`. The figure below demonstrates their magnitude spectrum. We observe that `Primitives` produces images that have a similar magnitude spectrum to those of natural images.



(a) (b) (c)

Figure 12. The magnitude spectrum of (a) Bridge, (b) `PinkNoise`, and (c) `Primitives`. We apply FFT on each image and then visualize the average magnitude of the images. When we visualize, we take a logarithmic transformation. Although `PinkNoise` aims to mimic the magnitude spectrum of natural images, that of `Primitives` approximates the benchmark dataset better than that of `PinkNoise`.

Table 4. KMMD score for **Table 3 in the main text (256)**.

	Obama	Cat	Brid.	Panda	Temp.	Wuzhen	Fountain	Build.
DfAug	0.23	0.15	0.23	0.28	0.18	0.39	0.21	0.21
TGAN	0.13	0.14	0.22	0.21	0.14	0.27	0.19	0.18
FrzD	**0.12**	**0.14**	0.22	**0.18**	**0.13**	0.25	0.21	**0.16**
Ours	0.17	0.15	**0.17**	0.26	0.14	**0.25**	**0.17**	0.18

11. Kernel Maximum Mean Discrepancy (KMMD)

Quantitative evaluation with various metrics is helpful to compare the models and understand the aspect. To this end, we also provide KMMD as suggested by Reviewer 1 in the rebuttal. We report FID only in the main text because of the following reason. In Figure 4(a) of [2], KMMD considers "*scale&shift*" as the best model although "*Ours*" provides more plausible results; "*scale&shift*" even failed to produce eye, nose, and mouth. Contrarily, FID ranked "*Ours*" as the best, correctly reflecting the perceptual fidelity. Table 4 shows the KMMD score of each model. Although the rankings with KMMD are slightly different from those with FID, our method similarly performs or outperforms the baselines. Overall, we conclude that `Primitives-PS` is still effective for pretraining GANs.

Table 5. FID score of *ImageNet pretrained* model and `Primitives-PS` pretrained model on 512×512.

	Obama	Cat	Brid.	Panda	Temp.	Wuzhen	Fountain	Build.
DfAug	59.6	28.0	147.8	14.4	45.0	150.9	214.2	99.2
TGAN	**37.5**	35.2	52.0	11.8	42.5	84.1	284.3	65.5
FrzD	39.1	28.8	**48.6**	**11.2**	**38.9**	**69.5**	**34.3**	**60.2**
Ours	50.8	**27.7**	51.6	14.9	41.9	81.6	42.9	80.9

12. Scale-up to higher resolution and comparison with ImageNet

To check the effectiveness of `Primitives-PS` in the higher resolution, we pretrain StyleGAN2 with `Primitives-PS` on 512×512, and then transfer to the low-shot datasets. Moreover, we use the ImageNet pretrained model for all competitors to investigate the effect of a diverse and large-scale training dataset. The pretrained file is from the link. We note that this model is pretrained on the 512×512 ImageNet until 1.3M steps. Since the ImageNet dataset can be considered as a super-set of eight test categories, the best performance using the ImageNet pretrained model is often better than `Primitives-PS` pretrained model. However, when the category of test set no longer overlaps with the ImageNet, we argue that only `Primitives-PS` can provide consistent and meaningful performances, *e.g., medical images for diagnoses, microscopic images for gene analysis or space imaging for navigation*. Besides, the pretrained model with the 1M ImageNet dataset is vulnerable to the private and copyright issue. A number of images contain a person and the copyright of each image might not be free to all the users. For these practical issues related to legality, the proposed `Primitives-PS` provides huge benefits for pretraining of GANs.

References

[1] 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. 11

[2] 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. 13

[3] 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. 10