Regularizing Generative Adversarial Networks under Limited Data Supplementary Materials

Hung-Yu Tseng*[2], Lu Jiang[1], Ce Liu[1], Ming-Hsuan Yang[1,2,4],
Weilong Yang[3]

[1]Google Research [2]University of California, Merced [3]Waymo [4]Yonsei University

1. Overview

In this supplementary document, we first provide the theoretical justification for Proposition 1 in the paper. Second, we describe the implementation details. Finally, we present additional experimental results, including those of training the GAN model on only hundreds of images, i.e., low-shot image generation.

2. Theoretical Justification

Proposition 1. *Consider the regularized objective in Eq.(1) and (2) in the paper for the WGAN [1], where R_{LC} is with a single anchor and $\lambda > 0$. Assume that with respect to a fixed generator G, the anchor converges to a stationary value α ($\alpha > 0$). Let $C(G)$ denote the generator's virtual objective for the fixed optimal D. We have:*

$$C(G) = (\frac{1}{2\lambda} - \alpha)\Delta(p_d \| p_g), \tag{1}$$

where $\Delta(P\|Q)$ is the LeCam-divergence aka the triangular discrimination [8] given by:

$$\Delta(P\|Q) = \sum_x \frac{(P(x) - Q(x))^2}{(P(x) + Q(x))} \tag{2}$$

Proof. In the following, we use $p_d(\boldsymbol{x})$ to denote the target distribution and simplify $\mathbb{E}_{\boldsymbol{z}\sim p_z(\boldsymbol{x})}\left[D(G(\boldsymbol{z}))\right]$ using $\mathbb{E}_{\boldsymbol{x}\sim p_g(\boldsymbol{x})}\left[D(\boldsymbol{x})\right]$. With a single anchor, the proposed regularization has the following form:

$$R_{\mathrm{LC}}(D) = \mathop{\mathbb{E}}_{\boldsymbol{x}\sim p_d(\boldsymbol{x})}\left[\|D(\boldsymbol{x}) + \alpha\|^2\right] + \mathop{\mathbb{E}}_{\boldsymbol{x}\sim p_g(\boldsymbol{x})}\left[\|D(\boldsymbol{x}) - \alpha\|^2\right], \tag{3}$$

where $\alpha \geq 0$ is the anchor for the real images, i.e., α_R in the Equation (4) in the paper. Note that since $D(G(\boldsymbol{z})) \leq 0$, when using a single anchor we have that $\alpha_R = -\alpha_F = \alpha$.

Consider the regularized objective of the discriminator:

$$\begin{aligned}
\min L(D) &= \min \mathop{\mathbb{E}}_{\boldsymbol{x}\sim p_g(\boldsymbol{x})}\left[D(\boldsymbol{x})\right] - \mathop{\mathbb{E}}_{\boldsymbol{x}\sim p_d(\boldsymbol{x})}\left[D(\boldsymbol{x})\right] + \lambda R_{\mathrm{LC}}(D) && (4)\\
&= \min \mathop{\mathbb{E}}_{\boldsymbol{x}\sim p_g(\boldsymbol{x})}\left[D(\boldsymbol{x})\right] - \mathop{\mathbb{E}}_{\boldsymbol{x}\sim p_d(\boldsymbol{x})}\left[D(\boldsymbol{x})\right] + \lambda \mathop{\mathbb{E}}_{\boldsymbol{x}\sim p_d(\boldsymbol{x})}\left[\|D(\boldsymbol{x}) + \alpha\|^2\right] + \lambda \mathop{\mathbb{E}}_{\boldsymbol{x}\sim p_g(\boldsymbol{x})}\left[\|D(\boldsymbol{x}) - \alpha\|^2\right] && (5)\\
&= \min \mathop{\mathbb{E}}_{\boldsymbol{x}\sim p_d(\boldsymbol{x})}\left[\lambda\|D(\boldsymbol{x}) + \alpha\|^2 - D(\boldsymbol{x})\right] + \mathop{\mathbb{E}}_{\boldsymbol{x}\sim p_g(\boldsymbol{x})}\left[\lambda\|D(\boldsymbol{x}) - \alpha\|^2 + D(\boldsymbol{x})\right] && (6)\\
&= \min \mathop{\mathbb{E}}_{\boldsymbol{x}\sim p_d(\boldsymbol{x})}\left[\lambda\|D(\boldsymbol{x}) + \alpha\|^2 - D(\boldsymbol{x}) - \alpha + \frac{1}{4\lambda}\right] + \mathop{\mathbb{E}}_{\boldsymbol{x}\sim p_g(\boldsymbol{x})}\left[\lambda\|D(\boldsymbol{x}) - \alpha\|^2 + D(\boldsymbol{x}) - \alpha + \frac{1}{4\lambda}\right] + C && (7)\\
&= \min \lambda \mathop{\mathbb{E}}_{\boldsymbol{x}\sim p_d(\boldsymbol{x})}\left[\|D(\boldsymbol{x}) + \alpha - \frac{1}{2\lambda}\|^2\right] + \lambda \mathop{\mathbb{E}}_{\boldsymbol{x}\sim p_g(\boldsymbol{x})}\left[\|D(\boldsymbol{x}) + \frac{1}{2\lambda} - \alpha\|^2\right] + C && (8)
\end{aligned}$$

*Work done during HY's internship at Google Research.

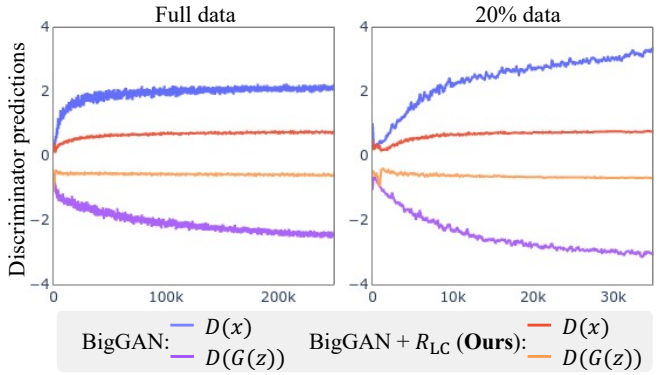


Figure 1: **Discriminator predictions.** We visualize the discriminator predictions from the BigGAN model on the CIFAR-10 dataset during the training stage. The proposed method prevents the predictions of real images $D(x)$ and generated images $D(G(z))$ from diverging under the limited (e.g., 20%) data setting.

where $C = 2\alpha - \frac{1}{2\lambda}$.

We now derive the optimal discriminator D^* with respect to a fixed G. According to the assumption, near convergence of D^*, C approaches a constant value. This is a mild assumption because we found that the discriminator predictions always converge to the stationary points in all of the experiments for both the WGAN and BigGAN models (cf. Fig. 8 in the main paper). Hypothetically speaking, in rare cases where this criterion might not hold, we may anneal the decay factor in the moving average α gradually to 1.0 near while D approaches convergence. In the following, we treat C as a constant value and compute D^* from:

$$D(\boldsymbol{x})^* = \operatorname*{argmin}_D L(D) = \lambda \int_{\boldsymbol{x}} \big[p_d(\boldsymbol{x})(D(\boldsymbol{x}) + \alpha - \frac{1}{2\lambda})^2 + p_g(\boldsymbol{x})(D(\boldsymbol{x}) - \alpha + \frac{1}{2\lambda})^2\big] dx \tag{9}$$

$$\frac{dL(D)}{dx} = 2\lambda\big[p_d(\boldsymbol{x})(D(\boldsymbol{x}) + \alpha - \frac{1}{2\lambda}) + p_g(\boldsymbol{x})(D(\boldsymbol{x}) - \alpha + \frac{1}{2\lambda})\big] = 0 \tag{10}$$

$$\implies (p_d(\boldsymbol{x}) + p_g(\boldsymbol{x}))D(\boldsymbol{x}) + (p_d(\boldsymbol{x}) - p_g(\boldsymbol{x}))(\alpha - \frac{1}{2\lambda}) = 0 \tag{11}$$

$$\implies D^*(\boldsymbol{x}) = \frac{(p_d(\boldsymbol{x}) - p_g(\boldsymbol{x}))(\frac{1}{2\lambda} - \alpha)}{p_d(\boldsymbol{x}) + p_g(\boldsymbol{x})} \tag{12}$$

Consider the following generator's objective when D is fixed:

$$\min_G L(G) = -\mathop{\mathbb{E}}_{\boldsymbol{x}\sim p_g(\boldsymbol{x})}\big[D(\boldsymbol{x})\big] + \mathop{\mathbb{E}}_{\boldsymbol{x}\sim p_d(\boldsymbol{x})}\big[D(\boldsymbol{x})\big] \tag{13}$$

Notice that as the regularization term is only added to the discriminator, and the generator's objective is kept the same. Then we have:

$$C(G) = \int_x \big[p_d(\boldsymbol{x})D^*(\boldsymbol{x}) - p_g(\boldsymbol{x})D^*(\boldsymbol{x})\big] dx \tag{14}$$

$$= (\frac{1}{2\lambda} - \alpha) \int_x \frac{(p_d(\boldsymbol{x}) - p_g(\boldsymbol{x}))^2}{p_d(\boldsymbol{x}) + p_g(\boldsymbol{x})} dx \tag{15}$$

$$= (\frac{1}{2\lambda} - \alpha)\Delta(p_d(\boldsymbol{x}) \| p_g(\boldsymbol{x})), \tag{16}$$

where Δ is the LeCam divergence and $\frac{1}{2\lambda} - \alpha$ is the weight of the divergence. Since the divergence is non-negative, we need $\lambda < \frac{1}{2\alpha}$. For example, if $\alpha = 1$, then $\lambda < 0.5$. This indicates the weight λ in the proposed regularization term R_{LC} should not be too large.

The proof is then completed. □

Discussion on the theoretical results. Our method is inspired by the theoretical analysis in Proposition 1. In our experiments, we employ modifications to optimize the performance. We note this is not a rare practice in the literature. For example,

Table 1: **Comparisons to WGAN on the CIFAR-10 dataset.** We report the average FID (↓) scores of three evaluation runs.

Methods	WGAN [1]		WGAN + R_{LC} (Ours)		BigGAN [2]		BigGAN + R_{LC} (Ours)	
	IS (↑)	FID (↓)	IS (↑)	FID (↓)	IS (↑)	FID (↓)	IS (↑)	FID (↓)
Full CIFAR-10	$7.86_{\pm.07}$	$18.86_{\pm.13}$	$7.98_{\pm.02}$	$15.79_{\pm.11}$	$9.07_{\pm0.03}$	$9.74_{\pm0.06}$	$\mathbf{9.31}_{\pm0.04}$	$\mathbf{8.31}_{\pm0.05}$

Goodfellow et al. [3] show theoretically the saturated GAN loss minimizes the JS-divergence. However, in practice, they use the non-saturated GAN due to the superior empirical performance.

Specifically, our method incorporates two modifications. First, it uses two anchors for the discriminator predictions of both real and generated images. Our ablation study in Table 6 in the main paper shows this leads to a performance gain. Second, we extend our method to regularize other GAN losses in the leading-performing GAN models such as the BigGAN [2] and StyleGAN2 [7] models. The former has a similar objective as the WGAN that applies the hinge loss [9]. Using a similar procedure in [1], we can show that the result in Proposition 1 also applies when the discriminator predictions are within the margin boundaries.

We empirically substantiate the analysis by showing the proposed regularization prevents the discriminator predictions from diverging on the limited training data. As shown in Figure 1, without regularization, the predictions of real and generated images diverge rapidly under the limited data setting. On the other hand, the proposed method keeps the predictions within -1 and +1.

3. Implementation Details

Exponential moving average. We implement the exponential moving average operation using the following formulation:

$$\alpha^{(t)} = \gamma \times \alpha^{(t-1)} + (1-\gamma) \times v^{(t)}, \tag{17}$$

where α is the moving average variable (i.e., α_R and α_F), $v^{(t)}$ is the current value at training step t, and γ is the decay factor. We set the decay factor γ to 0.99 in all experiments.

CIFAR-10 and CIFAR-100. We set the weight λ of regularization term to 0.3, and adopt the default hyper-parameters of the baseline method in the implementation by Zhao et al. [16].[1] Specifically, we use the batch size of 50, learning rate of $2e-4$ for the generator G and discriminator D, 4 D update steps per G step, and *translation* + *cutout* for the DA [16] method.

ImageNet. We use the Compare GAN codebase[2] for the experiments on the ImageNet dataset. The random scaling, random horizontal flipping operations are used to pre-process the images. We keep the default hyper-parameter settings for the baseline methods (i.e., BigGAN [2], BigGAN + DA [16]). As for our approach, we use the batch size of 2048, learning rate of $4e{-}4$ for D and $1e{-}4$ for G, 2 D update steps per G step, and the regularization weight λ of 0.01.

Comparisons with Data Augmentation We train and evaluate the StyleGAN2 [7] framework on the StyleGAN [6] dataset, wher is image size is 256×256. We set the regularization weight λ to $3e-7$ in this experiment. We use the ADA [5] codebase[3] and the DA [16] source code[4] for the experiments shown in Table 3 and Table 6 in the paper, respectively. Since the StyleGAN2 model uses the softplus mapping function for computing the GAN loss, the gradients of the discriminiator predictions around zero are much smaller than those in the BigGAN [2] model that uses the hinge function i.e., BigGAN: 0.8, StyleGAN2: 10^{-3} in the last layer of the discriminator). Therefore, we use a much smaller regularization weight λ of $3e-7$. Though the weight λ is smaller on the FFHQ dataset, we can observe the impact of our method by comparing StyleGAN2 (λ=0) and StyleGAN2+R_{LC} (λ=$3e-7$) in Table 3, 6 and 7 in the paper. Moreover, the ablation study results in Fig 7(b) suggest our approach is relatively insensitive to the value of λ under the same backbone. As for the other hyper-parameters, we keep the setting used in the original implementations.

Reproducing results of previous methods. We obtain quantitatively comparable results in most experiments. However, there are few cases that we fail to reproduce the results reported in the original paper. First, compared to Table 3 in the DA [16] paper, we obtain different results of training the StyleGAN2 model on the 5k and 1k FFHQ datasets, respectively. Second, the result of training the StyleGAN model with the ADA [5] method on the 1k FFHQ dataset is slightly different

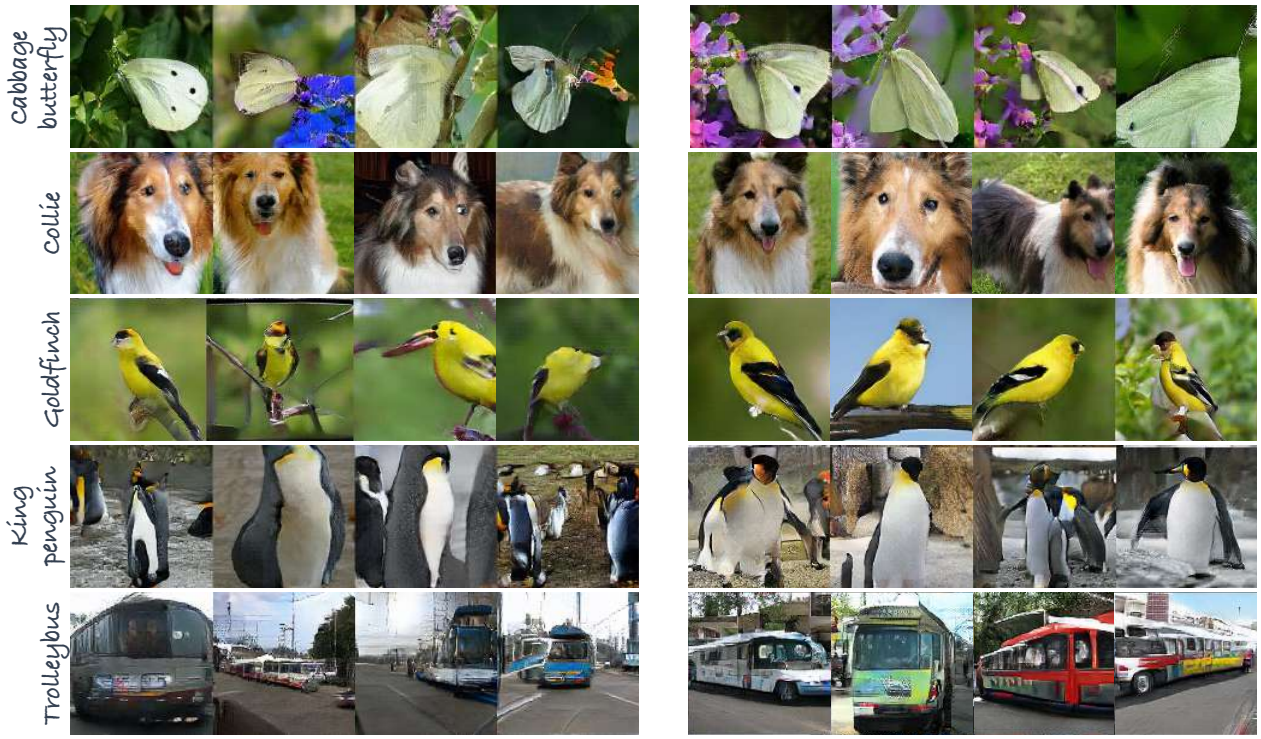
[1] https://github.com/mit-han-lab/data-efficient-gans/tree/master/DiffAugment-biggan-cifar
[2] https://github.com/google/compare_gan
[3] https://github.com/NVlabs/stylegan2-ada
[4] https://github.com/mit-han-lab/data-efficient-gans/tree/master/DiffAugment-stylegan2

Table 2: **IS scores on the CIFAR dataset.** We report the average IS scores (↑) of three evaluation runs to supplement Table 1 in the paper. The best performance is in **bold** and the second best is underscored.

Methods	CIFAR-10			CIFAR-100		
	Full data	20% data	10% data	Full data	20% data	10% data
Non-saturated GAN [3]	9.08±0.11	8.36±0.09	7.80±0.07	10.58±0.13	8.75±0.07	5.96±0.05
LS-GAN [10]	9.05±0.10	8.50±0.08	7.33±0.08	10.75±0.08	8.94±0.01	7.02±0.11
RaHinge GAN [4]	8.96±0.05	8.52±0.04	6.84±0.04	10.46±0.12	9.19±0.08	6.95±0.07
BigGAN [2]	9.07±0.03	8.52±0.10	7.09±0.03	10.71±0.14	8.58±0.04	6.74±0.04
BigGAN + R_{LC} (Ours)	**9.31**±0.04	**8.78**±0.07	**7.97**±0.03	**10.95**±0.07	**9.63**±0.06	**7.76**±0.01

Table 3: **Ablation study on exponential moving averages (EMAs).** We validate the impact of the EMAs by replacing the EMAs with the constant values. We train and evaluate the BigGAN model on the CIFAR dataset in this experiment.

FID(↓)	EMAs	[α_R=0.5, α_F=−0.5]	[α_R=1, α_F=−1]
CIFAR 10	**15.27**±0.10	30.64±0.05	19.81±0.03
CIFAR 100	**25.51**±0.19	30.03±0.11	27.54±0.07



BigGAN (FID: 15.01) BigGAN + R_{LC} **(Ours)** (FID: 11.16)

Figure 2: **Qualitative comparisons under limited training data.** We show the generation results on the 25% ImageNet dataset. The baseline models trained with our approach synthesize more realistic images.

from that reported in 7(c) in the ADA paper. On the other hand, the result of training the StyleGAN2 model on the full FFHQ dataset is similar to that shown in the DA and ADA papers. As a result, we argue that the different sets of limited data sampled for training the StyleGAN model (using the different random seeds) cause the performance discrepancy observed under the limited data setting.

Table 4: **Quatitative results on the low-shot image generation datasets.** We report the average FID scores (↓) of three evaluation runs. The best performance is **bold** and the second best is underscored. Using the proposed regularization approach along with data augmentation to train the model on only *100* (Obama, Grumpy cat, Panda), *160* (Cat), or *389* (Dog) images perform favorably against the transfer learning techniques that pre-train the model on *70000* images.

Methods	Pre-training?	100-shot [16]			AnimalFace [13]	
		Obama	Grumby cat	Panda	Cat	Dog
Scale/shift [12]	✓	50.72	34.20	21.38	54.83	83.04
MineGAN [14]	✓	50.63	34.54	14.84	54.45	93.03
TransferGAN [15]	✓	48.73	34.06	23.20	52.61	82.38
TransferGAN +DA [16]	✓	39.85	29.77	17.12	49.10	65.57
FreezeD [11]	✓	41.87	31.22	17.95	47.70	70.46
TransferGAN +DA	✓	35.75	29.32	14.50	46.07	61.03
StyleGAN2 [7]		80.45±.36	48.63±.05	34.07±.22	69.84±.19	129.9±.03
StyleGAN2 + DA		47.09±.14	27.21±.03	12.13±.07	42.40±.07	58.47±.06
StyleGAN2 + DA + R_{LC} (**Ours**)		**33.16**±.23	**24.93**±.12	**10.16**±.05	**34.18**±.11	**54.88**±.09

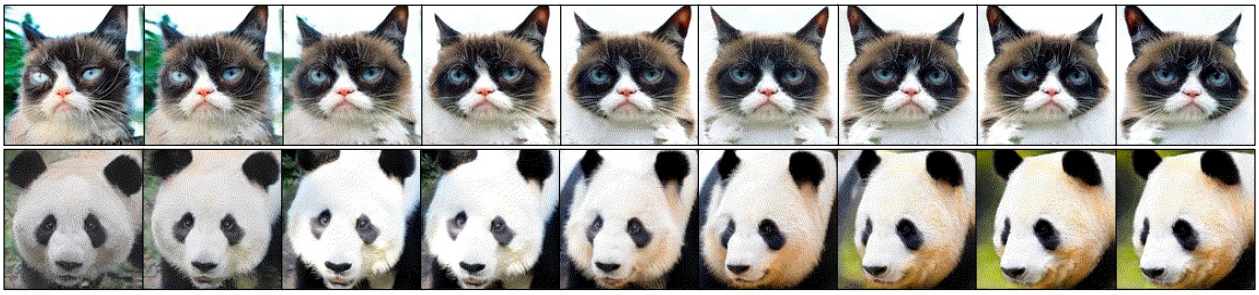


Figure 3: **Low-shot generation results.** We train the StyleGAN2 model with the proposed regularization and data augmentation methods on the 100 grumpy cat (*top*), Obama (*middle*), and panda (*bottom*) images.

4. Additional Experimental Results

4.1. CIFAR-10 and CIFAR-100

We report the results of WGAN on the CIFAR-10 dataset in Table 1. Although the proposed method is able to improve the performance of the WGAN model, the performance of the WGAN backbone is inferior to that of the BigGAN backbone and is also more sensitive to the hyperparameter setting. Therefore, we use the BigGAN backbone in our CIFAR and ImageNet experiments. In addition, Table 2 presents the IS scores to complement the FIS scores reported in Table 1 in the paper for the CIFAR experiments.

Necessity of exponential moving averages (EMAs). We validate the necessity of the EMAs in the table below with the BigGAN model on the 20% CIFAR datasets. Specifically, we compute our regularization with constant anchors by setting 1) α_R=1 and α_F=−1 following the LS-GAN [43] 2) α_F=−0.5 and α_R=0.5 (Figure 8 shows ± 0.5 is the converged value of EMAs.) The results in Table 3 show that using EMAs empirically facilitates the discriminator to converge to the better local optimal.

4.2. ImageNet

We show additional qualitative comparisons between the baseline (i.e., BigGAN [2]) and the proposed method (i.e., BigGAN + R_{LC}) in Figure 2. Combining the qualitative results shown in Figure 6 in the paper, we find that the proposed approach improves the visual quality of the generated images compared to the baseline models with and without data augmentation.

4.3. Low-Shot Image Generation

In this experiment, we consider a more extreme scenario where only a few dozens of images are available for training a GAN model. This setting is known as the *low-shot image generation* problem [16]. Recent solutions focus on adapting an exiting GAN model pre-trained on other large datasets. The adaptation strategies include optimizing the whole GAN

model [15], modifying the batch statics [12], using an additional mining network [14], and fine-tuning parts of the GAN model [11]. We use the experimental setting in the DA [16] paper that trains and evaluates the StyleGAN2 model on datasets that contain only 100 (Obama, Grumpy cat, Panda), 160 (Cat), or 389 (Dog) images.[4] We set the regularization weight λ to 0.0001. The quantitative comparisons are shown in Table 4. The StyleGAN2 model trained with the proposed regularization and data augmentation methods *from scratch* performs favorably against the existing adaptation-based techniques. Note that the adaptation-based approaches require to pre-train the StyleGAN2 model on the FFHQ dataset consisting of 70000 images. We also perform the interpolation in the latent space, and present the image generation results in Figure 3. More qualitative results are demonstrated in the supplementary video.

References

[1] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein gan. In *ICML*, 2017. 1, 3

[2] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. In *ICLR*, 2019. 3, 4, 5

[3] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. In *NIPS*, 2014. 3, 4

[4] Alexia Jolicoeur-Martineau. The relativistic discriminator: a key element missing from standard gan. In *ICLR*, 2019. 4

[5] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. *NeurIPS*, 2020. 3

[6] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, 2019. 3

[7] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *CVPR*, 2020. 3, 5

[8] Lucien Le Cam. *Asymptotic methods in statistical decision theory*. Springer Science & Business Media, 2012. 1

[9] Jae Hyun Lim and Jong Chul Ye. Geometric gan. *arXiv preprint arXiv:1705.02894*, 2017. 3

[10] Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. In *ICCV*, 2017. 4

[11] Sangwoo Mo, Minsu Cho, and Jinwoo Shin. Freeze discriminator: A simple baseline for fine-tuning gans. *arXiv preprint arXiv:2002.10964*, 2020. 5, 6

[12] Atsuhiro Noguchi and Tatsuya Harada. Image generation from small datasets via batch statistics adaptation. In *ICCV*, 2019. 5, 6

[13] Zhangzhang Si and Song-Chun Zhu. Learning hybrid image templates (hit) by information projection. *TPAMI*, 34(7):1354–1367, 2011. 5

[14] 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 *CVPR*, 2020. 5, 6

[15] Yaxing Wang, Chenshen Wu, Luis Herranz, Joost van de Weijer, Abel Gonzalez-Garcia, and Bogdan Raducanu. Transferring gans: generating images from limited data. In *ECCV*, 2018. 5, 6

[16] Shengyu Zhao, Zhijian Liu, Ji Lin, Jun-Yan Zhu, and Song Han. Differentiable augmentation for data-efficient gan training. *NeurIPS*, 2020. 3, 5, 6