Regularizing Generative Adversarial Networks under Limited Data

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

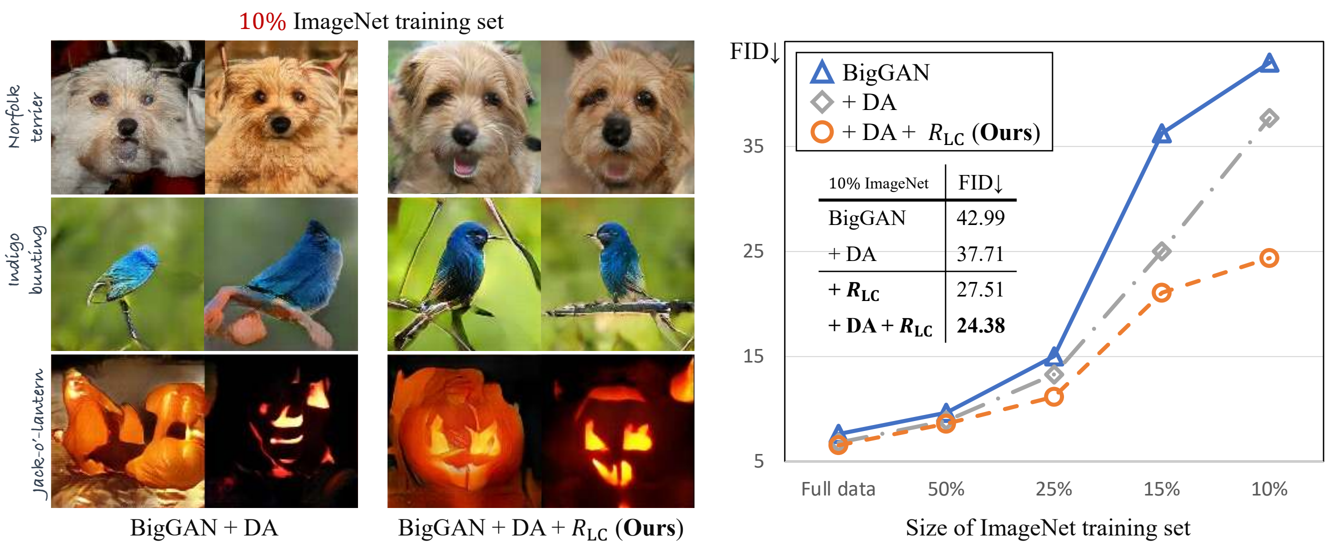

Figure 1: **Regularizing GANs under limited training data.** (*left*) Image generation trained on 10% ImageNet training set; (*right*) FID scores vs. ImageNet training set size. The proposed regularization method 1) addresses the limited training data issue for the GAN models, and 2) is empirically complementary to the recent data augmentation approaches [28, 71].

Abstract

Recent years have witnessed the rapid progress of generative adversarial networks (GANs). However, the success of the GAN models hinges on a large amount of training data. This work proposes a regularization approach for training robust GAN models on limited data. We theoretically show a connection between the regularized loss and an f-divergence called LeCam-divergence, which we find is more robust under limited training data. Extensive experiments on several benchmark datasets demonstrate that the proposed regularization scheme 1) improves the generalization performance and stabilizes the learning dynamics of GAN models under limited training data, and 2) complements the recent data augmentation methods. These properties facilitate training GAN models to achieve state-of-the-art performance when only limited training data of the ImageNet benchmark is available. The source code is available at `https://github.com/google/lecam-gan`.

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

1. Introduction

Generative adversarial networks (GANs) [2, 7, 13, 44] have made significant progress in recent years on synthesizing high-fidelity images. The GAN models are the cornerstone techniques for numerous vision applications, such as data augmentation [11, 12], domain adaptation [18, 19], image extrapolation [60], image-to-image translation [20, 34, 75], and image editing [1, 4, 21, 63].

The success of the GAN methods heavily relies on a large amount of diverse training data which is often labor-expensive or cumbersome to collect [65]. As the example of the BigGAN [7] model presented in Figure 1, the performance significantly deteriorates under the limited training data. Consequently, several very recent approaches [28, 71, 73] have been developed to address the data insufficiency issue. A representative task in this emerging research direction aims to learn a robust class-conditional GAN model when only a small proportion of the ImageNet data [54] are available for the training. Generally, existing methods exploit data augmentation, either conventional or differentiable augmentation, to increase the diversity of the limited training data. These data augmentation approaches have shown promising results on several standard benchmarks.

In this paper, we address the GAN training task on *limited data* from a different perspective: model regularization. Although there are numerous regularization techniques for the GAN models in the literature [14, 45, 47, 57, 74], none of them aim to improve the generalization of the GAN models trained on limited data. In contrast, our goal is to *learn robust GAN models on limited training data that can generalize well on out-of-sample data*. To this end, we introduce a novel regularization scheme to modulate the discriminator's prediction for learning a robust GAN model. Specifically, we impose an ℓ_2 norm between the current prediction of the real image and a moving average variable that tracks the historical predictions of the generated image, and vice versa. We theoretically show that, under mild assumptions, the regularization transforms the WGAN [2] formulation towards minimizing an f-divergence called LeCam-divergence [33]. We find that the LeCam-divergence is more robust under the limited training data setting.

We conduct extensive experiments to demonstrate the three merits of the proposed regularization scheme. First, it improves the generalization performance of various GAN approaches, such as BigGAN [7] and StyleGAN2 [29]. Second, it stabilizes the training dynamics of the GAN models under the limited training data setting. Finally, our regularization approach is empirically complementary to the data augmentation methods [28, 71]. As presented in Figure 1, we obtain state-of-the-art performance on the limited (e.g., 10%) ImageNet dataset by combining our regularization (i.e., R_{LC}) and the data augment method [71].

2. Related Work

Generative adversarial networks. Generative adversarial networks (GANs) [2, 7, 13, 25, 29, 44, 68] aim to model the target distribution using adversarial learning. Various adversarial losses have been proposed to stabilize the training or improve the convergence of the GAN models, mainly based on the idea of minimizing the f-divergence between the real and generated data distributions [50]. For example, Goodfellow et al. [13] propose the saturated loss that minimizes the JS-divergence between the two distributions. Similarly, the LSGAN [44] formulation leads to minimizing the χ^2-divergence [51], and the EBGAN [70] approach optimizes the total variation distance [2]. On the other hand, some models are designed to minimize the integral probability metrics (IPM) [48, 58], such as the WGAN [2, 14] frameworks. In this work, we design a new regularization scheme that can be applied to different GAN loss functions for training the GAN models on the limited data.

Learning GANs on limited training data. With the objective of reducing the data collection effort, several studies [15, 65] raise the concern of insufficient data for training the GAN models. Training the GAN models on limited data is challenging because the data scarcity leads to the problems such as unstable training dynamics, degraded fidelity of the generated images, and memorization of the training examples. To address these issues, recent methods [28, 62, 69, 71, 72, 73] exploit data augmentation as a mean to increase data diversity, hence preventing the GAN models from overfitting the training data. For example, Zhang et al. [69] augment the real images and introduce a consistency loss for training the discriminator. The DA [71] and ADA [28] approaches share a similar idea of applying differential data augmentation on both real and generated images, in which ADA further develops an adaptive strategy to adjust the probability of augmentation. In contrast to prior work, we tackle this problem from a different perspective of *model regularization*. We show that our method is conceptually and empirically complementary to the existing data augmentation approaches.

Regularization for GANs. Most existing regularization methods for GAN models aim to accomplish two goals: 1) stabilizing the training to ensure the convergence [45, 46], and 2) mitigating the mode-collapse issue [55]. As the GAN frameworks are known for unstable training dynamics, numerous efforts have been made to address the issue using noise [22, 57], gradient penalty [14, 30, 45, 53], spectral normalization [47], adversarial defense [74], etc. On the other hand, a variety of regularization approaches [5, 8, 41, 43, 59, 67] are proposed to alleviate the model-collapse issue, thus increasing the diversity of the generated images. Compared with these methods, our work targets a different goal: improving the generalization of the GAN models trained on the *limited training data*.

Robust Deep Learning. Robust deep learning aims to prevent the deep neural networks from overfitting or memorizing the training data. Recent methods have shown successes in overcoming training data bias such as label noise [3, 16, 23, 24, 52, 40, 49, 66] and biased data distributions [56, 10, 35, 38, 9]. Recently, few approaches [6, 26, 27, 61] have been proposed for learning the robust GAN model. While these approaches are designed to overcome label or image noise in a corrupted training set, we improve the generalization of the GAN models trained on the limited *uncorrupted* training data.

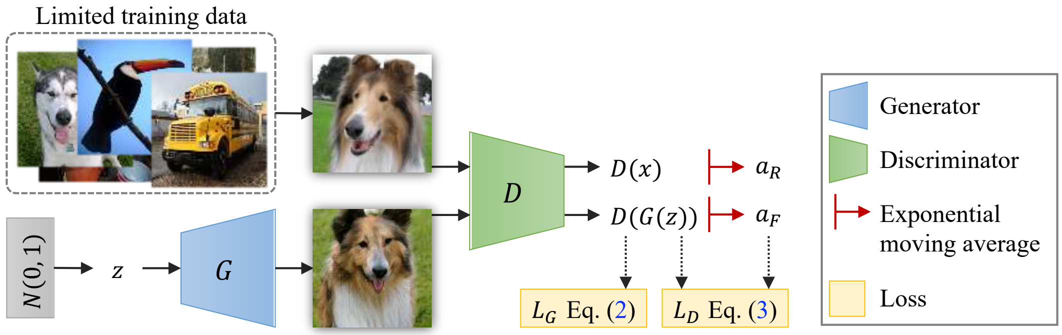

Figure 2: **Algorithmic overview.** During the GAN training stage, we use the exponential moving average variables, called *anchors*, to track the discriminator predictions. The anchors are then used to compute the regularized discriminator loss described in Eq. (3) to improve the generalization performance of the GAN models.

3. Methodology

We first review the GAN models, then detail our regularization scheme. Finally, we discuss the connection between the proposed method and the LeCam-divergence along with the effect on robust learning under the limited data setting.

3.1. Generative Adversarial Networks

A GAN model consists of a discriminator D and a generator G. Let V_D and L_G denote the training objectives of the discriminator D and generator G, respectively. The training of the GAN frameworks can be generally illustrated as:

$$\max_D V_D, V_D = \underset{\boldsymbol{x}\sim\mathcal{T}}{\mathbb{E}}\left[f_D(D(\boldsymbol{x}))\right] + \underset{\boldsymbol{z}\sim p_{\boldsymbol{z}}}{\mathbb{E}}\left[f_G(D(G(\boldsymbol{z})))\right] \quad (1)$$

$$\min_G L_G, L_G = \underset{\boldsymbol{z}\sim p_{\boldsymbol{z}}}{\mathbb{E}}\left[g_G(D(G(\boldsymbol{z})))\right], \quad (2)$$

where $p_{\mathbf{z}}$ is the prior distribution (e.g., $\mathcal{N}(0, I)$) and $\mathcal{T}$ is the training (observed) image set used to approximate the data distribution. The notations f_D, f_G, and g_G in Eq. (2) represent the mapping functions from which various GAN losses can be derived (cf. [39]).

3.2. Regularizing GANs under Limited Data

Our goal is to improve the performance of the GAN models when the training set $\mathcal{T}$ merely contains a limited amount of data, as the example shown in Figure 1. Different from the existing data augmentation methods [28, 71], we approach this problem by incorporating the regularization on the discriminator. We present the overview of the proposed method in Figure 2. The core idea is to regulate the discriminator predictions during the training phase. Specifically, we introduce two exponential moving average [32] variables α_R and α_F, called *anchors*, to track the discriminator's predictions of the real and generated images. The computation of the anchors α_R and α_F is provided in the supplementary document. We then use the identical objective L_G described in Eq. (2) for training the generator, and minimize the regularized objective L_D for the discriminator:

$$\min_D L_D, L_D = -V_D + \lambda R_{\mathrm{LC}}(D), \quad (3)$$

where R_{LC} is the proposed regularization term:

$$R_{\mathrm{LC}} = \underset{\boldsymbol{x}\sim\mathcal{T}}{\mathbb{E}}\left[\|D(\boldsymbol{x}) - \alpha_F\|^2\right] + \underset{\boldsymbol{z}\sim p_{\boldsymbol{z}}}{\mathbb{E}}\left[\|D(G(\boldsymbol{z})) - \alpha_R\|^2\right]. \quad (4)$$

At first glance, the objective in Eq. (3) appears counterintuitive since the regularization term R_{LC} pushes the discriminator to mix the predictions of real and generated images, as opposed to differentiating them. However, we show in Section 3.3 that R_{LC} offers meaningful constraints for optimizing a more robust objective. Moreover, we empirically demonstrate in Section 4 that with the appropriate weight λ, this simple regularization scheme 1) improves the generalization under limited training data, and 2) complements the existing data augmentation methods.

Why moving averages? Tracking the moving average of the prediction reduces the variance across mini-batches and stabilizes the regularization term described in Eq. (4). Intuitively, the moving average becomes stable while the discriminator's prediction gradually converges to the stationary point. We find this holds for the GAN models used in our experiments (e.g., Figure 8). We illustrate a general case of using two moving average variables α_R and α_F in Figure 2. In some cases, e.g., in theoretical analysis, we may use a single moving average variable to track the predictions of either real or generated images.

3.3. Connection to LeCam Divergence

We show the connection of the proposed regularization to the WGAN [2] model and an f-divergence called LeCam (LC)-divergence [33] or triangular discrimination [64]. Under mild assumptions, our regularization method can enforce WGANs to minimize the weighted LC-divergence. We show that the LC-divergence 1) can be used for training GAN models robustly under limited training data, and 2) has a close relationship with the f-divergences used in other GAN models [13, 44, 70].

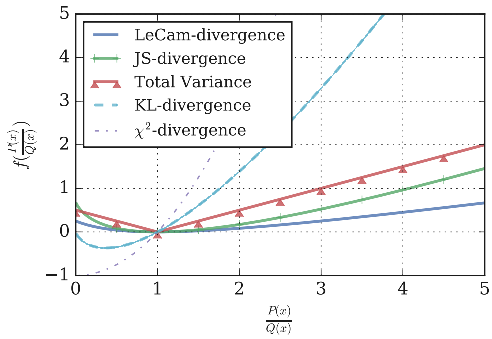

Figure 3: **Comparison of various f-divergences.** The x- and y-axis denote the input and the value of the function f in the f-divergence in Eq. (5). For extremely large or small inputs of $P(x)/Q(x)$, LeCam-divergence yields the most robust values of $f(P(x)/Q(x))$. The weighted LeCam-divergence is plotted where the weight is $\frac{1}{2\lambda} - \alpha = \frac{1}{4}$.

We first revisit the definition of the f-divergence. For two discrete distributions $Q(x)$ and $P(x)$, an f-divergence is defined as:

$$D_f(P\|Q) = \sum_x Q(x) f(\frac{P(x)}{Q(x)}) \tag{5}$$

if f is a convex function and $f(1) = 0$. The f-divergence plays a crucial role in GANs as it defines the underlying metric to align the generated distribution $p_g(\boldsymbol{x})$ and data distribution $p_d(\boldsymbol{x})$. For instance, Goodfellow et al. [13] showed that the saturated GAN minimizes the JS-divergence [37] between the two distributions:

$$C(G) = 2JS(p_d\|p_g) - \log(4), \tag{6}$$

where $C(G)$ is the virtual objective function for the generator when D is fixed to the optimal. Similarly, the LSGAN [44] method leads to minimizing the χ^2-divergence [51] and the EBGAN [70] scheme minimizes the total variation distance [2].

More recently, the Wasserstein distance [2], which does not belong to the f-divergence family, introduces a different distribution measurement. However, the performance of WGANs and similar models, e.g., BigGAN [7], deteriorates when the training data is limited. We show that incorporating the proposed regularization into these GAN models improves the generalization performance, especially under limited training data. Next, we show the connection between the regularized WGANs and LC-divergence.

Proposition 1. *Consider the regularized objective in Eq. (3) for the WGAN [2], 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 virtual objective function of the generator for the fixed optimal D. We have:*

$$C(G) = (\frac{1}{2\lambda} - \alpha)\Delta(p_d\|p_g), \tag{7}$$

where $\Delta(P\|Q)$ is the LeCam (LC)-divergence aka the triangular discrimination [33] given by:

$$\Delta(P\|Q) = \sum_x \frac{(P(x) - Q(x))^2}{(P(x) + Q(x))}. \tag{8}$$

Since the divergence is non-negative, we need $\lambda < \frac{1}{2\alpha}$, which indicates the regularization weight should not be too large. The proof is given in the supplementary materials. We note that the analysis in Proposition 1, which uses only a single anchor, is a simplified regularizer of our method described in Section 3.2. To achieve better performance and more general applications, we use 1) two anchors and 2) apply the regularization term to the hinge [7, 36] and non-saturated loss [13, 29] in the experiments. We note this is not a rare practice in the literature. For example, Goodfellow et al. [13] show theoretically the saturated GAN loss minimizes the JS-divergence. However, in practice, they use the non-saturated GAN for superior empirical results.

After drawing the connection between LC-divergence and regularized WGANs, we show that the LC-divergence is a robust f-divergence when limited data is available. Figure 3 illustrates several common f-divergences, where the x-axis plots the input to the function f in Eq. (5), i.e., $P(x)/Q(x)$, and the y-axis shows the function value of f. Note that the input $P(x)/Q(x)$ is expected to be erroneous when limited training data is available, and likely to include extremely large/small values. Figure 3 shows that the LC-divergence helps obtain a more robust function value for extreme inputs. In addition, the LC-divergence is symmetric and bounded between 0 and 2 which attains the minimum if and only if $p_d = p_g$. These properties demonstrate the LC-divergence as a robust measurement when limited training data is available. This observation is consistent with the experimental results shown in Section 4.

Proposition 2 (Properties of LeCam-divergence)**.** *LC-divergence Δ is an f-divergence with following properties:*

- *Δ is non-negative and symmetric.*
- *$\Delta(p_d\|p_g)$ is bounded, with the minimum 0 when $p_d = p_g$ and the maximum 2 when p_d and p_g are disjoint.*
- *Δ-divergence is a symmetric version of χ^2-divergence, i.e., $\Delta(P\|Q) = \chi^2(P\|M) + \chi^2(Q\|M)$, where $M = \frac{1}{2}(P+Q)$.*
- *The following inequalities hold [42]: $\frac{1}{4}\Delta(P,Q) \leq JS(P,Q) \leq \frac{1}{2}\Delta(P,Q) \leq \frac{1}{2}TV(P,Q)$, where JS and TV represent JS-divergence and Total Variation.*

Proposition 2 shows that the LC-divergence is closely related to the f-divergences used in other GAN methods. For example, it is a symmetric and smoothed χ^2-divergence used in the LSGAN [44]. The weighted Δ lower bounds the JS-divergence used in the saturated GAN [13] and the Total Variation distance used in the EBGAN [70] approaches.

Table 1: **Quantitative results on the CIFAR dataset.** We report the average FID scores (↓) of three evaluation runs. 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 [13]	$9.83_{\pm 0.06}$	$18.59_{\pm 0.15}$	$41.99_{\pm 0.18}$	$13.87_{\pm 0.08}$	$32.64_{\pm 0.19}$	$70.50_{\pm 0.38}$
LS-GAN [44]	$\underline{9.07}_{\pm 0.01}$	$\underline{21.60}_{\pm 0.11}$	$\underline{41.68}_{\pm 0.18}$	$\underline{12.43}_{\pm 0.11}$	$\mathbf{27.09}_{\pm 0.09}$	$54.69_{\pm 0.12}$
RaHinge GAN [25]	$11.31_{\pm 0.04}$	$23.90_{\pm 0.22}$	$48.13_{\pm 0.33}$	$14.61_{\pm 0.21}$	$\underline{28.79}_{\pm 0.17}$	$\underline{52.72}_{\pm 0.18}$
BigGAN [7]	$9.74_{\pm 0.06}$	$21.86_{\pm 0.29}$	$48.08_{\pm 0.10}$	$13.60_{\pm 0.07}$	$32.99_{\pm 0.24}$	$66.71_{\pm 0.01}$
BigGAN + R_{LC} (Ours)	$\mathbf{8.31}_{\pm 0.05}$	$\mathbf{15.27}_{\pm 0.10}$	$\mathbf{35.23}_{\pm 0.14}$	$\mathbf{11.88}_{\pm 0.12}$	$\mathbf{25.51}_{\pm 0.19}$	$\mathbf{49.63}_{\pm 0.16}$

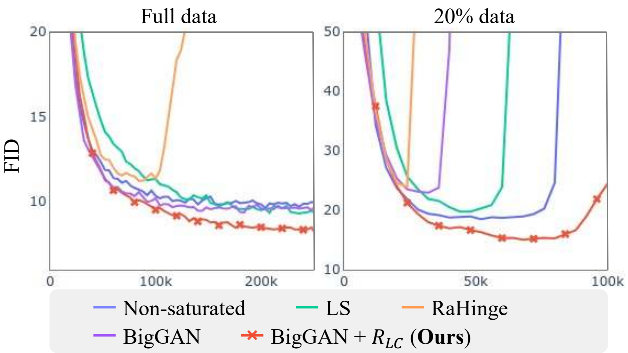

Figure 4: **FID curves during the training on the CIFAR-10 dataset.** The proposed method 1) improves the best performance, and 2) stabilizes the training dynamic of the BigGAN model under the limited (e.g., 20%) data setting.

Table 2: **Comparison to GAN regularization methods.** We report the average FID (↓) scores on the CIFAR datasets.

Method	CIFAR-10		CIFAR-100	
	Full data	20% data	Full data	20% data
BigGAN [7]	$9.74_{\pm .06}$	$21.86_{\pm .29}$	$13.60_{\pm .07}$	$32.99_{\pm .24}$
+ noise [57]	$9.64_{\pm .06}$	$21.87_{\pm .11}$	$13.88_{\pm .07}$	$32.38_{\pm .01}$
+ CR [69]	$8.96_{\pm .10}$	$20.62_{\pm .10}$	$\mathbf{11.59}_{\pm .05}$	$36.91_{\pm .12}$
+ GP-0 [45]	$10.30_{\pm .16}$	$19.10_{\pm .08}$	$14.67_{\pm .08}$	$29.85_{\pm .04}$
+ R_{LC} (Ours)	$\mathbf{8.31}_{\pm .05}$	$\mathbf{15.27}_{\pm .10}$	$11.88_{\pm .12}$	$\mathbf{25.51}_{\pm 0.19}$

4. Experimental Results

We conduct extensive experiments on several benchmarks to validate the efficacy of our method on training the leading class-conditional BigGAN [7] and unconditional StyleGAN2 [29] models on the limited data.

Datasets. The CIFAR 10/100 [31] and ImageNet [54] datasets are standard benchmarks for the image generation models. The resolutions of the images in the CIFAR, ImageNet datasets are 32x32, and 128x128, respectively.

Evaluation metrics. We use two common metrics: *Inception Score (IS)* [55] and *Fréchet Inception Distance (FID)* [17]. Unless specified otherwise, we follow the evaluation protocol in the DA paper [71] that reports the average and standard deviation values over three evaluation trials.

Setups. We conduct the CIFAR experiments using the BigGAN [7] framework implemented by Zhao et al. [71].[1] We train the BigGAN model on TPU for the ImageNet experiments.[2] Finally, the StyleGAN2 [29] framework is trained and evaluated using the implementation from Zhao et al. [71] and Karras et al. [28].[13] As for the hyper-parameter settings, we use the decay factor of 0.99 for the exponential moving average variables. We set the regularization weight λ to 0.3, 0.01 for the CIFAR, ImageNet experiments, respectively.

Baselines. We compare three types of baseline methods on the CIFAR datasets. The first group are GAN models that optimize various loss functions including ***non-saturated*** [13], ***LS*** [44], and ***RaHinge*** [25]. Second, we compare with three regularization methods: instance noise [57], zero-centered gradient penalty (***GP-0***) [45] and consistency regularization (***CR***) [69]. Finally, we compare with two recent differentiable data augmentation methods ***DA*** [71] and ***ADA*** [28] that address the limited data issue for GANs. For the experiments on other datasets, we focus on comparing with the state-of-the-art methods. For a fair comparison, we compare the baseline methods under the same GAN backbone using their official implementation on each dataset, except Table 5 in which we cite the numbers of [71] reported in the original paper.

4.1. Results on CIFAR-10 and CIFAR-100

As shown in Table 1, the proposed method improves the generalization performance of the BigGAN model. The comparison between other GAN models shows the competitive performance of the proposed method, especially under limited training data. These results substantiate that our regularization method minimizes a sensible divergence on limited training data. To further understand the impact on the training dynamics, we plot the FID scores during the training stage in Figure 4. The proposed method stabilizes the training process on limited data (i.e., FID scores deteriorate in a later stage) and achieves the lowest FID score at the final iteration (100K). This result suggests that our method can stabilize the GAN training process on limited data.

[1] https://github.com/mit-han-lab/data-efficient-gans
[2] https://github.com/google/compare_gan
[3] https://github.com/NVlabs/stylegan2-ada

Table 3: **Quantitative comparisons to data augmentation.** We report the average FID (↓) scores of three evaluation runs.

Methods	CIFAR-10		CIFAR-100		StyleGAN	
	Full	10%	Full	10%	Full	1K
BigGAN [7] + DA [71]	8.75±0.05	23.34±0.28	11.99±0.10	35.39±0.16	-	-
BigGAN + DA + R_{LC} (Ours)	**8.46**±0.06	**16.69**±0.02	**11.20**±0.09	**27.28**±0.05	-	-
StyleGAN2 [29] + ADA [28]	2.68±0.02	6.72±0.03	3.04±0.02	14.06±0.07	3.82±0.01	23.27±0.14
StyleGAN2 + ADA+ R_{LC} (Ours)	**2.47**±0.01	**6.56**±0.02	**2.99**±0.01	**13.01**±0.02	**3.49**±0.04	**21.70**±0.06

Table 4: **Quantitative results on the ImageNet dataset**. We report the mean IS (↑) and FID (↓) scores of three training runs.

Methods	Full data		50% data		25% data	
	IS ↑	FID ↓	IS ↑	FID ↓	IS ↑	FID ↓
BigGAN [7]	90.48±12.7	8.60±1.08	80.26±5.55	9.83±0.94	61.05±6.43	18.22±2.59
BigGAN + R_{LC} (Ours)	**93.00**±3.27	**7.27**±0.14	**89.94**±6.67	**9.13**±0.84	**65.66**±4.96	**14.47**±1.73

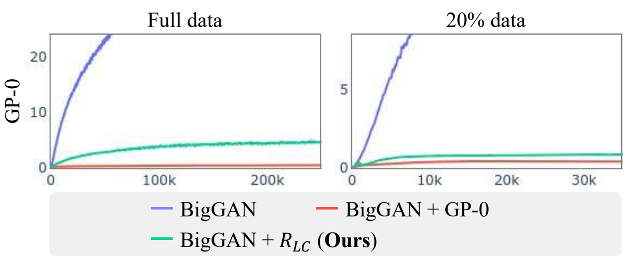

Figure 5: **Zero-centered gradient penalty values.** We visualize the values of zero-centered gradient penalty (GP-0) [45] during the training stage. The proposed regularization also constrains the values without explicitly minimizing the GP-0 loss.

We compare our method with three regularization methods: instance noise [57], GP-0 [45] and CR [69] in Table 2. Notice that the spectral norm regularization [47] is used by default in the BigGAN model [7]. For the GP-0 method, we apply the gradient penalty only on real images. Our regularization scheme performs favorably against these regularization methods, particularly under the limited data setting. Despite the improvement under the limited data, the GP-0 approach degrades the FID performance when using the full training data. We note that a similar observation is raised in the BigGAN paper [7]. In Figure 5, we visualize the GP-0 values of the models trained with the GP-0 and our methods during the training stage. Interestingly, the proposed method also constrains the GP-0 values, although it does not explicitly minimize the GP-0 loss.

Finally, we combine our regularization method with data augmentation and show it is complementary to the recent data augmentation methods [28, 71]. As presented in Table 3, the proposed approach improves the performance of DA and ADA, especially under the limited data settings. Note that the data augmentation methods tackle the problem from different perspectives and represent the prior state-of-the-art on limited training data before this work.

4.2. Comparison to State-of-the-art on ImageNet

ImageNet [54] is a challenging dataset since it contains more categories and images with higher resolution. Considering the variance of the model performance, we follow the evaluation protocol in the BigGAN paper [7]. Specifically, we run the training/evaluation pipeline three times using different random seeds, then report the average performance. We present the quantitative results in Table 4. The proposed method improves the resistance of the BigGAN model against the scarce training data issue (e.g., ↓ 3.75 in FID under 25% data). It is noteworthy that the performance variance of our models is reduced in most cases (e.g., 2.59 → 1.73 in FID under 25% data), suggesting its capability in stabilizing the training process.

Table 5 demonstrates the quantitative results compared to the state-of-the-art model that uses the DA [71] method. Both the quantitative results and qualitative comparison presented in Figure 6 validate that the proposed method complements the data augmentation approach. We achieve state-of-the-art performance on the limited (e.g., 10%) ImageNet dataset by combining our regularization and the data augment approaches.

4.3. Comparison with Data Augmentation

We use the StyleGAN dataset to conduct the experiments.Experiment details are provided in the supplementary document. As presented in Table 3 and Table 6, the proposed method improves the performance of the StyleGAN2 model trained with(out) data augmentation [28] in all cases. We note that different from BigGAN, the StyleGAN2 model minimizes the non-saturated [13] GAN loss and uses the gradient penalty GP-0 [45] in the default setting. This shows that the proposed regularization scheme can be applied to other GAN loss functions along with existing regularization approaches.

We make a comparison in Table 7 to summarize the (dis)advantages of the data augmentation and our methods. First, the data augmentation approaches yield more

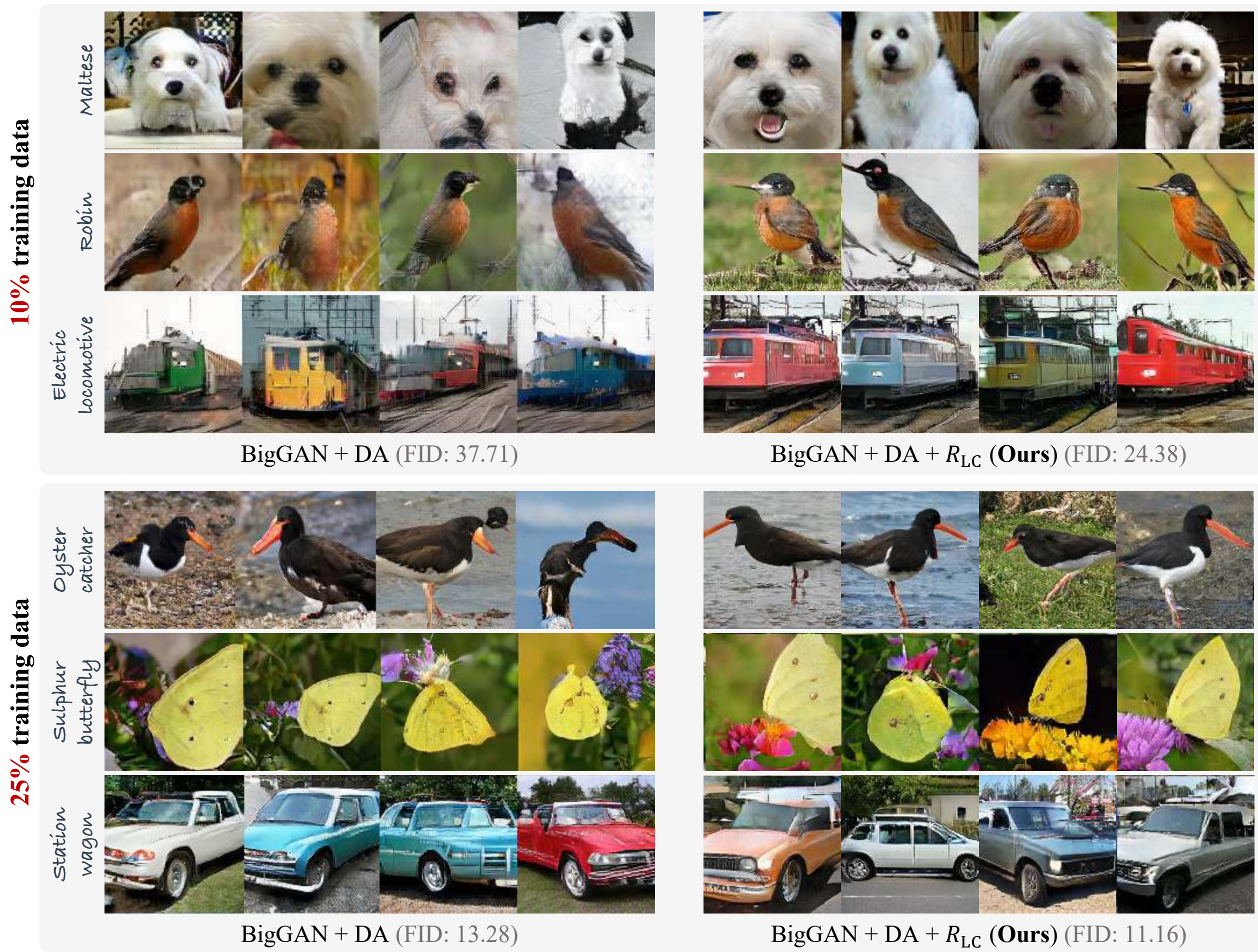

Figure 6: **Qualitative comparisons under limited training data.** We show the generation results on the (*top*) 10% and (*bottom*) 25% ImageNet dataset. The baseline models trained with our approach synthesize more realistic images.

Table 5: **Comparison to the state-of-the-art on the limited ImageNet training data.**. We train and evaluate the BigGAN [7] model following the same evaluation protocol in [71]. † denotes the result is quoted from [71].

Methods	Full data		50% data		25% data		10% data	
	IS ↑	FID ↓	IS ↑	FID ↓	IS ↑	FID ↓	IS ↑	FID ↓
DA [71] (Zhao et al.)	$100.8_{\pm 0.2}$†	$6.80_{\pm 0.02}$†	$\mathbf{91.9}_{\pm 0.5}$†	$8.88_{\pm 0.06}$†	$74.2_{\pm 0.5}$†	$13.28_{\pm 0.07}$†	$27.7_{\pm 0.1}$	$37.71_{\pm 0.11}$
DA + R_{LC} (Ours)	$\mathbf{108.0}_{\pm 0.6}$	$\mathbf{6.54}_{\pm 0.03}$	$91.7_{\pm 0.6}$	$\mathbf{8.59}_{\pm 0.01}$	$\mathbf{84.7}_{\pm 0.5}$	$\mathbf{11.16}_{\pm 0.05}$	$\mathbf{42.3}_{\pm 0.3}$	$\mathbf{24.38}_{\pm 0.06}$

significant gain than the proposed method when the training data is extremely limited. Nevertheless, our method can further improve the performance of data augmentation due to the complementary nature of the two methods. Second, the data augmentation approaches may degrade the performance when the training images are sufficiently diverse (e.g., the full dataset). This is consistent with the observation described in [28]. In comparison, our regularization method may not suffer the same problem.

4.4. Analysis and Ablation Studies

We use the BigGAN model and the CIFAR-10 dataset to conduct the analysis and ablation studies.

Regularization strength for R_{LC}. We conduct a sensitive study on the regularization weight λ. As shown in Figure 7(b), weights greater than 0.5 degrade the performance. This agrees with our analysis in Eq. (7) that larger weights λ result in negative divergence values. Generally, the proposed method is effective when the weight λ is in a reasonable range, e.g., $[0.1, 0.5]$ in Figure 7(b).

Regularizing real vs. generated image predictions. Our default method regularizes the predictions of both real images $D(x)$ and generated images $D(G(z))$. In this experiment, we investigate the effectiveness of separately regularizing the two terms $D(x)$ and $D(G(z))$. As shown in

Table 6: **Quantitative results of the StyleGAN2 [29] model.** We report the average FID (↓) scores of three evaluation runs.

Method	70k images	30k images	10k images	5k images	1k images
StyleGAN2 [29]	$3.79_{\pm 0.02}$	$6.19_{\pm 0.05}$	$14.96_{\pm 0.05}$	$25.88_{\pm 0.09}$	$72.07_{\pm 0.04}$
StyleGAN2 + R_{LC} (Ours)	$\mathbf{3.66}_{\pm 0.02}$	$\mathbf{5.78}_{\pm 0.03}$	$\mathbf{14.58}_{\pm 0.04}$	$\mathbf{23.83}_{\pm 0.11}$	$\mathbf{63.16}_{\pm 0.11}$

Table 7: **Comparisons with data augmentation methods.** We report the FID (↓) scores of the StyleGAN2 backbone.

Method	Full data	1K data
StyleGAN2 [29]	$3.71_{\pm 0.01}$	$72.07_{\pm 0.04}$
+ DA [71]	$4.21_{\pm 0.03}$	$25.17_{\pm 0.09}$
+ ADA [28]	$3.81_{\pm 0.01}$	$23.27_{\pm 0.14}$
$+R_{\text{LC}}$	$3.66_{\pm 0.02}$	$63.16_{\pm 0.11}$
+ ADA + R_{LC}	$\mathbf{3.49}_{\pm 0.04}$	$\mathbf{21.70}_{\pm 0.06}$

Table 8: **Ablation study on regularizing real vs. generated image predictions.** We train and evaluate the BigGAN [7] model on the CIFAR-10 dataset, then report the average FID (↓) scores.

Real	Generated	Full data	20% data
		$9.74_{\pm 0.06}$	$21.86_{\pm 0.29}$
✓		$8.73_{\pm 0.04}$	$20.47_{\pm 0.36}$
	✓	$8.79_{\pm 0.09}$	$18.18_{\pm 0.08}$
✓	✓	$\mathbf{8.31}_{\pm 0.03}$	$\mathbf{15.27}_{\pm 0.10}$

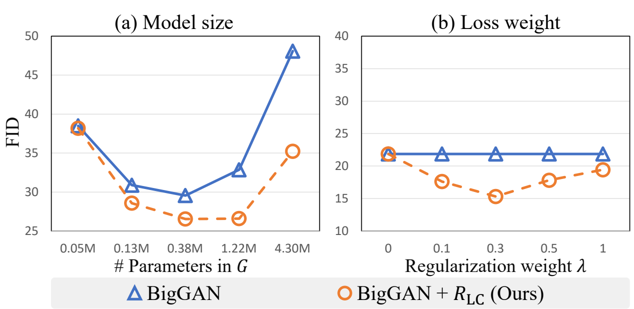

Figure 7: **Different (a) model sizes and (b) regularization strengths.** The scores are computed on the (a) 10% and (b) 20% CIFAR-10 datasets.

Table 8, regularizing both terms achieves the best result.

Discriminator predictions. We visualize the discriminator predictions during training in Figure 8. Without regularization, the predictions of real and generated images diverge rapidly as the discriminator overfits the limited training data. On the other hand, the proposed method, as described in Eq. (4), penalizes the difference between predictions of real and generated images, thus keeping the predictions in a particular range. This observation empirically substantiates that the discriminator's prediction gradually converges to the stationary point, and so do the moving average variables α_R and α_F.

Model size. Since reducing the model capacity may alleviate the overfitting problem, we investigate the performance of using a smaller model size for both generator and discriminator. Figure 7(a) shows the results of progressively halving the number of channels in both the generator and discriminator. The improvement made by our method increases with the model size, as the overfitting issue is more severe for the model with higher capacity.

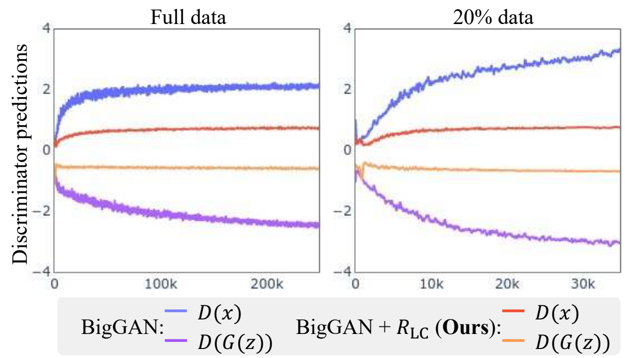

Figure 8: **Discriminator predictions.** We visualize the discriminator predictions from the BigGAN model on the CIFAR-10 dataset during the training stage. The proposed method prevents the predictions of real images $D(x)$ and generated images $D(G(z))$ from diverging under the limited (e.g., 20%) data setting.

5. Conclusion and Future Work

In this work, we present a regularization method to train the GAN models under the limited data setting. The proposed method achieves a more robust training objective for the GAN models by imposing a regularization loss to the discriminator during the training stage. In the experiments, we conduct experiments on various image generation datasets with different GAN backbones to demonstrate the efficacy of the proposed scheme that 1) improves the performance of the GAN models, especially under the limited data setting and 2) can be applied along with the data augmentation methods to further enhance the performance. In future, we plan the training data scarcity issue for 1) the conditional GAN tasks such as image extrapolation, image-to-image translation, etc, and 2) the robust GAN learning on large-scale noisy training data.

Acknowledgements

This work is supported in part by the NSF CAREER Grant #1149783.

References

[1] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan++: How to edit the embedded images? In *CVPR*, 2020. 1

[2] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein gan. In *ICML*, 2017. 1, 2, 3, 4

[3] Devansh Arpit, Stanislaw Jastrzkebski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, et al. A closer look at memorization in deep networks. In *ICML*, 2017. 2

[4] David Bau, Hendrik Strobelt, William Peebles, Jonas Wulff, Bolei Zhou, Jun-Yan Zhu, and Antonio Torralba. Semantic photo manipulation with a generative image prior. *ACM TOG (Proc. SIGGRAPH)*, 38(4), 2019. 1

[5] David Berthelot, Thomas Schumm, and Luke Metz. Began: Boundary equilibrium generative adversarial networks. *arXiv preprint arXiv:1703.10717*, 2017. 2

[6] Ashish Bora, Eric Price, and Alexandros G Dimakis. Ambientgan: Generative models from lossy measurements. In *ICLR*, 2018. 2

[7] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. In *ICLR*, 2019. 1, 2, 4, 5, 6, 7, 8

[8] Tong Che, Yanran Li, Athul Paul Jacob, Yoshua Bengio, and Wenjie Li. Mode regularized generative adversarial networks. In *ICLR*, 2017. 2

[9] Yong Cheng, Lu Jiang, Wolfgang Macherey, and Jacob Eisenstein. Advaug: Robust adversarial augmentation for neural machine translation. In *ACL*, 2020. 2

[10] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *CVPR*, 2019. 2

[11] 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. 1

[12] Maayan Frid-Adar, Eyal Klang, Michal Amitai, Jacob Goldberger, and Hayit Greenspan. Synthetic data augmentation using gan for improved liver lesion classification. In *International Symposium on Biomedical Imaging*, 2018. 1

[13] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. In *NeurIPS*, 2014. 1, 2, 3, 4, 5, 6

[14] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. In *NeurIPS*, 2017. 2

[15] Ishaan Gulrajani, Colin Raffel, and Luke Metz. Towards gan benchmarks which require generalization. In *ICLR*, 2019. 2

[16] Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. Coteaching: Robust training of deep neural networks with extremely noisy labels. In *NeurIPS*, 2018. 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. In *NeurIPS*, 2017. 5

[18] 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 *ICML*, 2018. 1

[19] Han-Kai Hsu, Chun-Han Yao, Yi-Hsuan Tsai, Wei-Chih Hung, Hung-Yu Tseng, Maneesh Singh, and Ming-Hsuan Yang. Progressive domain adaptation for object detection. In *WACV*, 2020. 1

[20] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *ECCV*, 2018. 1

[21] Minyoung Huh, Richard Zhang, Jun-Yan Zhu, Sylvain Paris, and Aaron Hertzmann. Transforming and projecting images into class-conditional generative networks. In *ECCV*, 2020. 1

[22] Simon Jenni and Paolo Favaro. On stabilizing generative adversarial training with noise. In *CVPR*, 2019. 2

[23] Lu Jiang, Mason Liu Di Huang, and Weilong Yang. Beyond synthetic noise: Deep learning on controlled noisy labels. In *ICML*, 2020. 2

[24] Lu Jiang, Zhengyuan Zhou, Thomas Leung, Li-Jia Li, and Li Fei-Fei. Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels. In *ICML*, 2018. 2

[25] Alexia Jolicoeur-Martineau. The relativistic discriminator: a key element missing from standard gan. In *ICLR*, 2019. 2, 5

[26] Takuhiro Kaneko and Tatsuya Harada. Noise robust generative adversarial networks. In *CVPR*, 2020. 2

[27] Takuhiro Kaneko, Yoshitaka Ushiku, and Tatsuya Harada. Label-noise robust generative adversarial networks. In *CVPR*, 2019. 2

[28] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. *NeurIPS*, 2020. 1, 2, 3, 5, 6, 7, 8

[29] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *CVPR*, 2020. 2, 4, 5, 6, 8

[30] Naveen Kodali, Jacob Abernethy, James Hays, and Zsolt Kira. On convergence and stability of gans. *arXiv preprint arXiv:1705.07215*, 2017. 2

[31] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. *Technical report*, 2009. 5

[32] Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. In *ICLR*, 2017. 3

[33] Lucien Le Cam. *Asymptotic methods in statistical decision theory*. Springer Science & Business Media, 2012. 2, 3, 4

[34] Hsin-Ying Lee, Hung-Yu Tseng, Qi Mao, Jia-Bin Huang, Yu-Ding Lu, Maneesh Singh, and Ming-Hsuan Yang. Drit++: Diverse image-to-image translation via disentangled representations. *IJCV*, pages 1–16, 2020. 1

[35] Junwei Liang, Lu Jiang, and Alexander Hauptmann. Simaug: Learning robust representations from simulation for trajectory prediction. In *ECCV*, 2020. 2

[36] Jae Hyun Lim and Jong Chul Ye. Geometric gan. *arXiv preprint arXiv:1705.02894*, 2017. 4

[37] Jianhua Lin. Divergence measures based on the shannon entropy. *IEEE Transactions on Information Theory*, 37(1):145–151, 1991. 4

[38] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *ICCV*, 2017. 2

[39] Ming-Yu Liu, Xun Huang, Jiahui Yu, Ting-Chun Wang, and Arun Mallya. Generative adversarial networks for image and video synthesis: Algorithms and applications. *arXiv preprint arXiv:2008.02793*, 2020. 3

[40] Sheng Liu, Jonathan Niles-Weed, Narges Razavian, and Carlos Fernandez-Granda. Early-learning regularization prevents memorization of noisy labels. In *NeurIPS*, 2020. 2

[41] Shaohui Liu, Xiao Zhang, Jianqiao Wangni, and Jianbo Shi. Normalized diversification. In *CVPR*, 2019. 2

[42] Guoxiang Lu and Bingqing Li. A class of new metrics based on triangular discrimination. *Information*, 6(3):361–374, 2015. 4

[43] Qi Mao, Hsin-Ying Lee, Hung-Yu Tseng, Siwei Ma, and Ming-Hsuan Yang. Mode seeking generative adversarial networks for diverse image synthesis. In *CVPR*, 2019. 2

[44] Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. In *ICCV*, 2017. 1, 2, 3, 4, 5

[45] Lars Mescheder, Andreas Geiger, and Sebastian Nowozin. Which training methods for gans do actually converge? In *ICML*, 2018. 2, 5, 6

[46] Lars Mescheder, Sebastian Nowozin, and Andreas Geiger. The numerics of gans. In *NeurIPS*, 2017. 2

[47] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. In *ICLR*, 2018. 2, 6

[48] Alfred Müller. Integral probability metrics and their generating classes of functions. *Advances in Applied Probability*, pages 429–443, 1997. 2

[49] Curtis G Northcutt, Lu Jiang, and Isaac L Chuang. Confident learning: Estimating uncertainty in dataset labels. *Journal of Artificial Intelligence Research*, 2021. 2

[50] Sebastian Nowozin, Botond Cseke, and Ryota Tomioka. f-gan: Training generative neural samplers using variational divergence minimization. In *NeurIPS*, 2016. 2

[51] Karl Pearson. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 50(302):157–175, 1900. 2, 4

[52] Mengye Ren, Wenyuan Zeng, Bin Yang, and Raquel Urtasun. Learning to reweight examples for robust deep learning. In *ICML*, 2018. 2

[53] Kevin Roth, Aurelien Lucchi, Sebastian Nowozin, and Thomas Hofmann. Stabilizing training of generative adversarial networks through regularization. In *NeurIPS*, 2017. 2

[54] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *IJCV*, 115(3):211–252, 2015. 2, 5, 6

[55] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *NeurIPS*, 2016. 2, 5

[56] Jun Shu, Qi Xie, Lixuan Yi, Qian Zhao, Sanping Zhou, Zongben Xu, and Deyu Meng. Meta-weight-net: Learning an explicit mapping for sample weighting. In *NeurIPS*, 2019. 2

[57] Casper Kaae Sønderby, Jose Caballero, Lucas Theis, Wenzhe Shi, and Ferenc Huszár. Amortised map inference for image super-resolution. *arXiv preprint arXiv:1610.04490*, 2016. 2, 5, 6

[58] Jiaming Song and Stefano Ermon. Bridging the gap between f-gans and wasserstein gans. In *ICML*, 2020. 2

[59] Akash Srivastava, Lazar Valkov, Chris Russell, Michael U Gutmann, and Charles Sutton. Veegan: Reducing mode collapse in gans using implicit variational learning. In *NeurIPS*, 2017. 2

[60] Piotr Teterwak, Aaron Sarna, Dilip Krishnan, Aaron Maschinot, David Belanger, Ce Liu, and William T Freeman. Boundless: Generative adversarial networks for image extension. In *ICCV*, 2019. 1

[61] Kiran Koshy Thekumparampil, Ashish Khetan, Zinan Lin, and Sewoong Oh. Robustness of conditional gans to noisy labels. In *NeurIPS*, 2018. 2

[62] Ngoc-Trung Tran, Viet-Hung Tran, Ngoc-Bao Nguyen, Trung-Kien Nguyen, and Ngai-Man Cheung. Towards good practices for data augmentation in gan training. *arXiv preprint arXiv:2006.05338*, 2020. 2

[63] Hung-Yu Tseng, Matthew Fisher, Jingwan Lu, Yijun Li, Vladimir Kim, and Ming-Hsuan Yang. Modeling artistic workflows for image generation and editing. In *ECCV*, 2020. 1

[64] István Vincze. On the concept and measure of information contained in an observation. In *Contributions to Probability*, pages 207–214. Elsevier, 1981. 3

[65] Ryan Webster, Julien Rabin, Loïc Simon, and Frédéric Jurie. Detecting overfitting of deep generative networks via latent recovery. In *CVPR*, 2019. 1, 2

[66] Youjiang Xu, Linchao Zhu, Lu Jiang, and Yi Yang. Faster meta update strategy for noise-robust deep learning. In *CVPR*, 2021. 2

[67] Dingdong Yang, Seunghoon Hong, Yunseok Jang, Tianchen Zhao, and Honglak Lee. Diversity-sensitive conditional generative adversarial networks. In *ICLR*, 2019. 2

[68] Han Zhang, Ian J Goodfellow, Dimitris N Metaxas, and Augustus Odena. Self-attention generative adversarial networks. In *ICML*, 2019. 2

[69] Han Zhang, Zizhao Zhang, Augustus Odena, and Honglak Lee. Consistency regularization for generative adversarial networks. In *ICLR*, 2020. 2, 5, 6

[70] Junbo Zhao, Michael Mathieu, and Yann LeCun. Energy-based generative adversarial network. In *ICLR*, 2017. 2, 3, 4

[71] Shengyu Zhao, Zhijian Liu, Ji Lin, Jun-Yan Zhu, and Song Han. Differentiable augmentation for data-efficient gan training. *NeurIPS*, 2020. 1, 2, 3, 5, 6, 7, 8

[72] Zhengli Zhao, Sameer Singh, Honglak Lee, Zizhao Zhang, Augustus Odena, and Han Zhang. Improved consistency regularization for gans. *arXiv preprint arXiv:2002.04724*, 2020. 2

[73] Zhengli Zhao, Zizhao Zhang, Ting Chen, Sameer Singh, and Han Zhang. Image augmentations for gan training. *arXiv preprint arXiv:2006.02595*, 2020. 2

[74] Brady Zhou and Philipp Krähenbühl. Don't let your discriminator be fooled. In *ICLR*, 2019. 2

[75] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, 2017. 1