CONSISTENCY REGULARIZATION FOR GENERATIVE ADVERSARIAL NETWORKS

Anonymous authors
Paper under double-blind review

ABSTRACT

Generative Adversarial Networks are plagued by training instability, despite considerable research effort. Progress has been made on this topic, but many of the proposed interventions are complicated, computationally expensive, or both. In this work, we propose a simple and effective training stabilizer based on the notion of Consistency Regularization - a popular technique in the Semi-Supervised Learning literature. In particular, we augment data passing into the GAN discriminator and penalize the sensitivity of the ultimate layer of the discriminator to these augmentations. This regularization reduces memorization of the training data and demonstrably increases the robustness of the discriminator to input perturbations. We conduct a series of ablation studies to demonstrate that the consistency regularization is compatible with various GAN architectures and loss functions. Moreover, the proposed simple regularization can consistently improve these different GANs variants significantly. Finally, we show that applying consistency regularization to GANs improves state-of-the-art FID scores from 14.73 to 11.67 on the CIFAR-10 dataset.

1 INTRODUCTION

Generative Adversarial Networks (GANs) (Goodfellow et al., 2014) are a powerful framework for generative modeling that have achieved especially impressive results on image synthesis benchmarks(Brock et al., 2018; Salimans et al., 2016; Zhang et al., 2018; Miyato et al., 2018; Miyato & Koyama, 2018). In the original setting, a GAN is composed of two neural networks trained with competing goals. The *generator* is trained to synthesize realistic samples to fool the discriminator, whereas the *discriminator* is trained to distinguish real samples from fake ones produced by the generator. This training process corresponds to a minimax two-player game.

One major problem with GANs is the instability of the training procedure and the general sensitivity of the results to various hyperparameters(Salimans et al., 2016). Because GAN training implicitly requires finding the Nash equilibrium of a non-convex game in a continuous and high dimensional parameter space, it is substantially more complicated than standard neural network training. In fact, formally characterizing the convergence properties of the GAN training procedure is mostly an open problem (Odena, 2019). Previous work (Miyato et al., 2018; Odena et al., 2017) has provided evidence that interventions focused on the discriminator are helpful for mitigating some of these stability issues. (Miyato et al., 2018) propose a technique called spectral normalization, in which weight matrices in the discriminator are divided by an approximation of their largest singular value. (Arjovsky et al., 2017) and (Gulrajani et al., 2017) propose forms of gradient norm penalties, in which the norm of the discriminator jacobian is indirectly penalized. DRAGAN (Kodali et al., 2017) introduces another form of gradient penalty where the gradients at Gaussian perturbations of training data are penalized. All of these forms of regularization introduce non-trivial run-time overheads (Kurach et al., 2019b) and considerably implementation complexity.

In the interest of simplifying things and reducing computational cost, we looked to a technique called Consistency Regularization (Sajjadi et al., 2016; Zhai et al., 2019; Ye et al., 2017; Xie et al., 2019). Roughly speaking, the over-arching theme of existing GAN 'regularizers' is that they enforce that the discriminator' outputs do not change too much with respect to small changes in the input. Consistency regularization asks for something similar, but not exactly the same: that the outputs of the model under consideration do not change too much with respect to transformations that *should* be

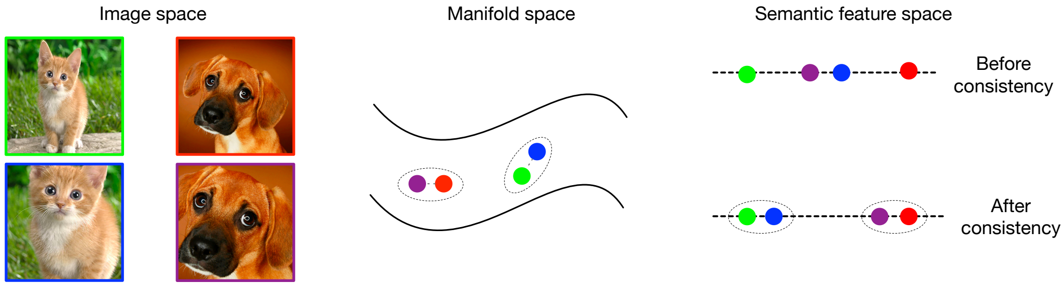

Figure 1: A stylized illustration of consistency regularization for GANs. Before consistency regularization, a GAN discriminator might 'think' that the sideways dog and the sideways cat are more similar to each other than they are to their respective augmentations (i.e. zoom-in in our example), by virtue of them being sideways. This is illustrated in the upper right, where the purple dot is closer to the blue dot than to the green dot, and so forth. After we enforce consistency regularization based on the implicit assumption that image augmentation preserves the semantics we care about, the purple dot pulled closer to the red dot.

semantics preserving. For example, we might flip images horizontally and ask an object recognition model to yield the same result for both the image and its flipped version. The difference between this technique and, say, spectral normalization is that spectral normalization doesn't have any notion of what a semantics preserving transformation is. In light of the similarity between consistency regularization and GAN 'regularization techniques', it is natural to wonder if consistency regularization can be applied to GAN discriminators. In this work, we introduce a technique to do just that: we augment images before they are fed into the GAN discriminator and penalize the sensitivity of the ultimate layer of the discriminator to those augmentations.

This technique is simple, but surprisingly effective. It is much less computationally expensive than prior techniques, and much less complicated to implement and test. In extensive ablation studies, we show that it works across a large range of GAN variants and data sets. We also show that simply applying this technique on top of existing GAN models yields new state-of-the-art results as measured by Frechet Inception Distance (Heusel et al., 2017).

In summary, our contributions are:

- We apply consistency regularization to GAN discriminators to yield a simple, effective regularizer with substantially lower computational cost than gradient-based regularization methods.
- We conduct a variety of experiments with different data sets and GAN variants to demonstrate that our technique is generally applicable.
- We conduct experiments showing that applying consistency regularization to GANs can yield substantial improvements over prior state of the art in image synthesis. For example, the proposed consistency regularization improves state-of-the-art FID scores from 14.73 to 11.67 on the CIFAR-10 dataset.

2 Method

2.1 GANs

A GAN consists of a generator network and a discriminator network. The generator $\mathcal{G}(z)$ takes a latent variable $z \sim q(z)$ from a prior distribution and maps it deterministically to the observation space X. The discriminator $\mathcal{D}(x)$ takes an observation $x \in X$ and produces a posterior distribution over possible observation sources (either G or the target distribution). In the standard GAN training procedure we attempt to solve the following min-max optimization problem:

$$\min_{\mathcal{G}} \max_{\mathcal{D}} \ \mathbb{E}_{x \sim \mathbb{P}_r} \left[L(\mathcal{D}(x), 1) \right] + \mathbb{E}_{x \sim q(z)} \left[L(\mathcal{D}(\mathcal{G}(z)), 0) \right] \tag{1}$$

where $q(z)$ is usually a standard normal distribution and $L(\cdot, \cdot)$ is a loss function that enforces that $\mathcal{D}$ output 1 if its input is from the target distribution and 0 if its input is from the generator. In practice, this means doing simultaneous gradient descent on the parameters of G and D.

A significant amount of research has been done on modifying this formulation in order to somehow improve the training results. A notable example is Arjovsky et al. (2017), in which the authors propose clipping the weights of the discriminator in an attempt to enforce that the GAN training procedure implicitly optimizes a bound on the Wassertein distance between the target distribution and the distribution given by the generator. Subsequent work has refined this technique in several steps (Gulrajani et al., 2017; Miyato et al., 2018; Zhang et al., 2018), and the current 'best practice' is to enforce spectral normalization on both the generator and the discriminator.

2.2 Consistency Regularization

Consistency regularization has emerged as the gold-standard (Zhai et al., 2019; Sajjadi et al., 2016; Ye et al., 2017; Xie et al., 2019; Oliver et al., 2018) for Semi-Supervised Learning on image data. The basic idea is straightforward: an input image is perturbed in some semantics preserving way and the sensitivity of the classifier to that perturbation is penalized. The perturbation can take many forms: it can be image flipping, or cropping, or adversarial 'attacks'. The precise form of the regularization can also take many forms: sometimes it is a mean-squared-error on model logits and sometimes it is some form of divergence between the distribution over classes implied by the logits. All forms of consistency regularization can be seen as a way to use unlabeled data to find a sub-manifold of observation space $\mathcal{X}$ that our image data set is 'close' to (Belkin et al., 2006). Of course, this is only useful if we believe the hypothesis that the data actually do lie close to some low dimensional manifold. For the rest of this paper we will assume that this hypothesis – often called the Manifold Assumption or the Manifold Hypothesis (Chapelle et al., 2009) – is basically true for modern image data sets.

2.3 Consistency Regularization for GANs

If the Manifold Hypothesis is mostly true, how can this help us train GANs? Informally speaking, we'd like to 'tell the generator' to assume the Manifold Hypothesis. Since the generator parameters come from performing gradient descent on the discriminator activations, enforcing that the discriminator roughly obeys the manifold hypothesis seems like a good way to accomplish this.

Thus, we propose performing consistency regularization on the GAN discriminator during training. In practice, we randomly augment training images as they are passed to the discriminator and penalize the sensitivity of the discriminator's ultimate layer to those augmentations.

We write $\mathcal{D}_j(x)$ to mean the output before activation of the jth layer of the discriminator given output x. We write $T(x,t)$ to denote a stochastic data augmentation function indexed by time t, so that $T(x,t)$ can be read as 'the result of the augmentation applied to image x at time t'. This function can be linear or nonlinear, but is understood to preserve the semantics[1] of the input. Our proposed regularization is defined as:

$$\min_{\mathcal{D}} \mathcal{L}_{cs} = \min_{\mathcal{D}} \sum_{j=m}^{n} \big(\mathcal{D}_j(x) - \mathcal{D}_j(T(x,t))\big)^2, \tag{2}$$

where m denote m-th layer and totally n layers in the discriminator architecture. This consistency regularization enforces the discriminator produce the same features for the data point x under different data augmentations. Empirically, high-level features are usually learned in the later layers of the discriminator, while low-level and middle-level features are learned in the earlier layers.

In our experiments, we find that consistency regularization on the last layer of the discriminator before activation is sufficient, so L_{cs} can be specifically rewritten as:

$$L_{cs} = \big(D_n(x) - D_n(T(x,t))\big)^2, \tag{3}$$

where from now on we will drop the j for brevity. Therefore the overall Consistency-Regularized GAN objective is defined as

[1] Where semantics in this case is understood to mean 'whether the image should be classified as coming from the target distribution or from the Generator.'

Algorithm 1 Consistency Regularization Batch Update

Input: Parameters θ_G, θ_D, weight λ, learning rate α
Output: Updated parameters
$z \sim p(z)$ ▷ Draw batch from prior
$x \sim p_{\text{data}}(x)$ ▷ Draw batch from target distribution
$L_{cs} \leftarrow (D(x) - D(T(x,t)))^2$ ▷ Compute consistency loss
$\theta_G \leftarrow \theta_G - \alpha \nabla_{\theta_G}(L_G(D(G(z))))$ ▷ Update G as normal
$\theta_D \leftarrow \theta_D - \alpha \nabla_{\theta_D}(L_D(D(G(z), D(x))) + \lambda L_{cs})$ ▷ Update D using added consistency loss

$$\min_{\mathcal{G}} \max_{\mathcal{D}} \; \mathbb{E}_{x\sim\mathbb{P}_r}\big[L(\mathcal{D}(x),1)\big] + \mathbb{E}_{x\sim q(z)}\big[L(\mathcal{D}(\mathcal{G}(z)),0)\big] + \mathbb{E}_{x\sim\mathbb{P}_r}\big[\,(\mathcal{D}_n(x) - \mathcal{D}_n(T(x,t)))^2\,\big]. \tag{4}$$

This cost is just added to the discriminator loss (weighted by a hyper-parameter λ) when updating the discriminator parameters. The generator update remains unchanged. See Algorithm 1 for more details. Note that the only extra computational cost due to this method comes from feeding a third more images through the discriminator forward and backward when updating the discriminator parameters. We speculate that this cost could be reduced further by throwing away half of the training examples used to update the generator, augmenting only the remaining training examples, and passing a minibatch to the discriminator of the same size as the one passed to the generator, but we have not tried this.

3 Experiments

This section explores the empirical results of applying consistency regularization to GANs. First, we compare our consistency-regularized GAN models with state-of-the-art GAN models (Miyato & Koyama, 2018; Zhang et al., 2018; Brock et al., 2018). And then, the proposed consistency 'regularizer' is compared with previous GAN regularization techniques(Kodali et al., 2017; Gulrajani et al., 2017) for several GAN architectures and loss functions. Finally, we investigate the proposed consistency regularization with extensive ablation study.

All our experiments are based on the open-source codebase of compare GAN (Kurach et al., 2019a), available at https://github.com/google/compare_gan.

3.1 Datasets and Evaluation Metrics

We test our proposed method on two datasets: CIFAR-10 (Krizhevsky, 2009) and CELEBA-HQ-128 (Karras et al., 2018). We follow the procedure in (Kurach et al., 2019b) to prepare datasets. CIFAR-10 consists of 60 thousand of 32×32 images in 10 classes; 50 thousand for training and 10 thousand for testing. CELEBA-HQ-128 (CELEBA) contains 30k images of faces at a resolution of 128×128. We use three thousand of images for testing and the rest of images for training.

We adopt the Fréchet Inception distance (FID) (Heusel et al., 2017) for quantitative evaluation. FID directly estimates the distance between the target distribution and the distribution implied by the trained generator. It is broadly thought to be consistent with human evaluation in assessing the realism and variation of the generated samples (Heusel et al., 2017). In our experiments the FID score is calculated on the test dataset. By default, the best FID for each model is reported.

3.2 Compare with state-of-the-art GAN models

We first compare our proposed consistency-regularized GAN models with the state-of-the-art (Miyato & Koyama, 2018; Brock et al., 2018) for class conditional image generation on CIFAR-10. As shown in Table 1, our models achieve the best FID, with the proposed consistency regularization. Our model renews the state-of-the-art FID on CIFAR-10 from 14.73 to 11.67.

To investigate whether directly adding the proposed consistency regularization can further boost the state-of-the art GAN models. Our model has exactly the same architecture and is trained under

Dataset	SNGAN (Miyato & Koyama, 2018)	BigGAN (Brock et al., 2018)	BigGAN*	Ours
CIFAR-10	17.5	14.73	20.42	11.67

Table 1: Comparison of the proposed consistency-regularized GAN models with state-of-the-art GAN models (Miyato & Koyama, 2018; Brock et al., 2018) for class conditional image generation on CIFAR-10 in terms of FID. BigGAN* is the BigGAN implementation is the compare gan codebase. Our model has the exactly same architecture as BigGAN* and trained under the same setting, with the only difference that our model has the proposed consistency regularization.

the same setting as BigGAN*, the open sourced implementation of BigGAN in the compare GAN codebase. The only difference is that our model has the proposed consistency regularization. As reported in Table 1, the FID score of BigGAN on CIFAR-10 is improved from 20.42 to 11.67, which demonstrates that the proposed consistency regularization can significantly improve state-of-the-art for image generation.

We will explain more details about why the consistency regularization works in the ablation studies section.

3.3 Comparison with other GAN 'Regularizers'

We compare our methods with state-of-the-art GAN regularization techniques(Kodali et al., 2017; Gulrajani et al., 2017) on CIFAR-10 and CELEBA datasets. We follow the procedure from (Kurach et al., 2019a) to evaluate across several GAN architectures and loss functions. Specifically, we use a ResNet (He et al., 2016; Gulrajani et al., 2017) architecture for CIFAR-10, and a SNDCGAN (Miyato et al., 2018) model for CELEBA. For CIFAR-10, we evaluate regularization methods using three loss functions:

- the non-saturating loss (NS) (Goodfellow et al., 2014)
- the Wasserstein loss (Wass) (Arjovsky et al., 2017)
- the hinge loss (Hinge) (Lim & Ye, 2017; Tran et al., 2017)

On CELEBA, we used the hinge loss for simplicity. All models are trained in the compare GAN codebase with its default settings. Result for each model is based on three runs with different random seeds. The top ten best FID scores for each run are collected for drawing Figure 2. We use mini-batches of size 64 and we stop training after 200 thousand discriminator update steps for CIFAR-10 and 100 thousand steps for for CELEBA.

The results are illustrated in Figure 2. Our consistency-regularized models achieve the lowest (best) mean FID for all settings across different GAN architectures, loss functions and datasets. For example, our models significantly outperform all baseline models by adding the consistency regularization. Especially, even when the baseline models completely fail to converge, *e.g.,* settings (e)(f) of Figure 2, reasonable FID scores are achieved by our consistency-regularized models. The results indicate that the proposed consistency regularization is stable enough to be used as a plug-and-play technique to improve GAN performance in different settings without much tuning.

In contrary, previous regularization techniques, such as DRAGAN(Kodali et al., 2017) and Gradient Penalty (GP)(Gulrajani et al., 2017), are more difficult to be plugged in the baseline model to improve its performance without tuning the architecture or hyper-parameters. As shown in Figure 2 (a)(b)(c)(e), under the default setting of compare GAN codebase, adding DRAGAN or GP regularization could not improve the baseline models. In some setting, *e.g.,* (f)(g) of Figure 2, although the baseline models are significant improved by DRAGAN and GP regularizers, they can perform even better with our consistency regularization. Consequently, the comparison result further demonstrates the advantage of the proposed consistency regularization. This finding is especially encouraging, considering that the proposed consistency regularization has substantially lower computational cost (and is substantially simpler to implement) than the other techniques. In our experiments, the consistency regularization is generally around 1.5 to 1.8 times faster than gradient based regularization techniques, such as DRAGAN and GP, which need to compute higher order gradients of the discriminator function.

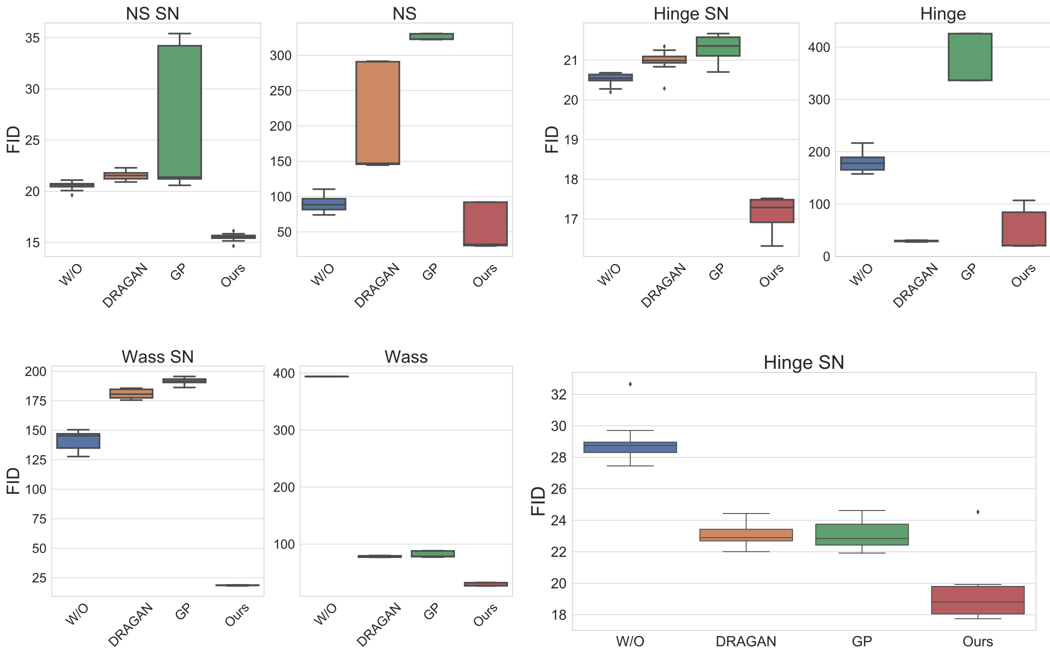

Figure 2: Comparison of our proposed consistency regularizer with previous regularization techniques, baseline model without regularization (W/O), DRAGAN(Kodali et al., 2017), and Gradient Penalty (GP)(Gulrajani et al., 2017), on different hyper-parameter and loss settings on CIFAR-10 (a-f) and CELEBA (g) datasets. (a) spectral normalization (SN) with non-saturating (NS) loss. (b) NS w/o SN. (c) hinge loss (Hinge) with SN. (d) Hinge w/o SN. (e) Wasserstein loss (Wass) with SN. (f) Wass w/o SN. (g) Hinge with SN.

4 Ablation Studies and Discussion

4.1 How Much Does Augmentation Matter by Itself?

Our consistency regularization technique actually has two parts: we perform data augmentation on inputs from the training data, and then we perform consistency regularization on both the augmented data and the original data. But how much of the effect we measure in Section 3 are due to data augmentation alone? After all, it's well known that data augmentation can reduce over-fitting for object recognition models, and a GAN discriminator is an object recognition model.

An experiment is designed to answer this question. First, we train three GANs:

- A GAN trained using Algorithm 1.
- A baseline GAN, trained without augmentation or Consistency Regularization.
- A GAN trained with only data augmentation, and no Consistency Regularization.

And then, we plot both their FID and the accuracy of the discriminator on a held-out test set in Figure 3. The FID tells us how 'good' the resulting GAN is, and the discriminator accuracy tells us how much the GAN discriminator over-fits.

Interestingly, we found that these two measures were not well correlated in this case. The model trained with only data augmentation over-fits substantially less than the baseline GAN, but has almost exactly the same FID. The model trained with the consistency regularization has the same amount of over-fitting as the model trained with just data augmentation, but a much lower FID. This suggests an interesting hypothesis, which is that the mechanism by which the consistency regularization improves GANs is not simply discriminator generalization. The robustness to the data perturbation is more likely to be main reason for the impressive gains from consistency regularization.

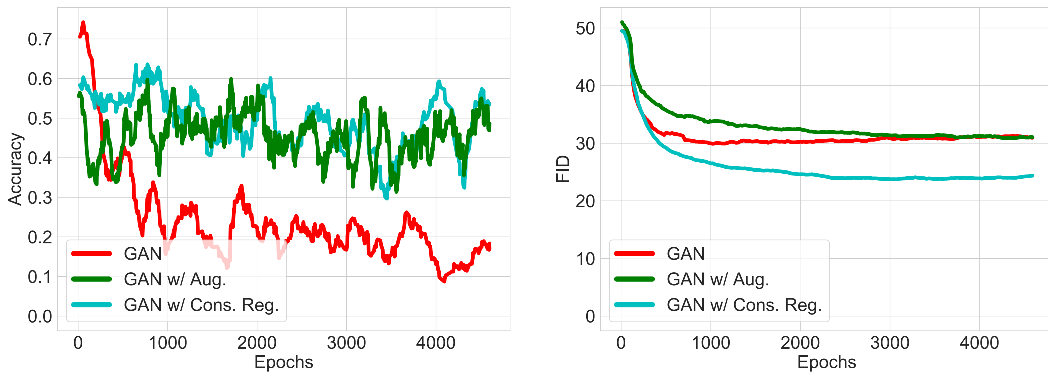

Figure 3: A study of how much data augmentation matters by itself. Three GANs were trained on CIFAR-10: one baseline GAN, one GAN with data augmentation only, and one GAN with both data augmentation and consistency regularization. (**Left**) Accuracy of the GAN discriminator on the held out test set. The accuracy is low for the baseline GAN, which indicates it suffered from over-fitting. The accuracy for the other two basically indistinguishable for each other. This suggests that augmentation by itself is enough to reduce discriminator over-fitting, and that consistency regularization by itself does little to address over-fitting. (**Right**) FID scores. The score for the GAN with only augmentation is not any better than the score for the baseline, even though its discriminator is not over-fitting. The score for the GAN with consistency regularization is better than both of the others, suggesting that the consistency regularization acts on the score through some mechanism other than by reducing discriminator over-fitting.

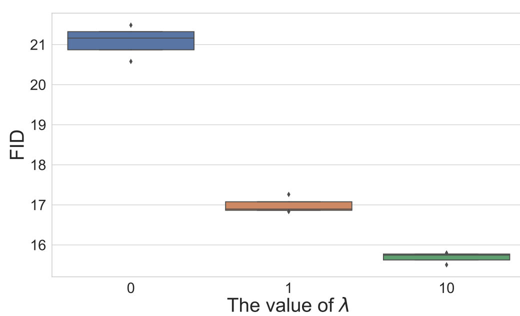

Figure 4: CIFAR-10 GANs with various λ values. All the other hyper-parameters are held constant. A value of 100 sometimes causes training to diverge. Otherwise, the FID of the generated samples improves monotonically as we increase λ.

4.2 How does the Type of Augmentation Affect Results?

To analyze how different types of data augmentation affect our results, we conduction an ablation study on the CIFAR-10 dataset comparing the results of using four different types of image augmentation:

- Adding Gaussian noise to the image in pixel-space
- Randomly shifting the image by a few pixels and then flipping it horizontally.
- Applying cutout (DeVries & Taylor, 2017) transformations to the image.
- Cutout *and* random shifting and flipping.

As shown in Table 2, random flipping and shifting *without* cutout gives the best results (FID 16.04, Inception Score 8.28) among all four methods. Adding Gaussian noise in pixel-space gives the worst results. This could be because adding Gaussian noise is inconsistent with the Manifold Hypothesis.

It's also noteworthy that the most extensive augmentation (random flipping and shifting with cutout) did not perform the best - this suggests that it can be useful to start with simple augmentations and

Metric	Gaussian Noise	Random shift	Cutout	Cutout w/ random shift
FID	21.91±0.32	16.04±0.17	17.10±0.29	19.46±0.26
Inception Score	8.12±0.02	8.28±0.02	8.18±0.40	8.16±0.10

Table 2: Scores of GAN quality on CIFAR-10 for different types of image augmentation. Gaussian noise is the worst, and Random Shift is the best, consistent with general consensus on the best way to perform image optimization on CIFAR-10 (Zagoruyko & Komodakis, 2016). Interestingly, the most substantial augmentation does not yield the best performance.

gradually increase their complexity until results plateau. Notably, random flipping and shifting has been adopted as the de-facto standard data augmentation policy on the CIFAR-10 dataset (Zagoruyko & Komodakis, 2016), which is consistent with our results. For simplicity and because it performs the best, we use random shifting and flipping as the default data augmentation policy for all of our other experiments.

4.3 How does λ Affect Results?

λ is the weight of the consistency loss in the overall discriminator loss function. To better understand the effect of λ, we train GANs on CIFAR-10 for λ values of $0, 1, 10$, and 100 while keeping all the other hyperparameters fixed. A value of 100 causes training to diverge roughly a third of the time. Otherwise – as shown in Figure 4 – the FID of the generated samples improves monotonically as we increase λ. For this reason, we fixed $\lambda = 10$ in all our experiments.

5 Conclusion

In this paper, we have applied consistency regularization to GANs. This is a simple step, but a lot of progress in the GAN literature has come from similar simple steps (Radford et al., 2016; Denton et al., 2015; Odena et al., 2017; Zhang et al., 2018; Brock et al., 2018). Consistency regularization is extremely simple to implement, computationally cheap, and results in improvements in all of the many contexts in which we evaluated it. We also conducted a thorough study on hyper-parameters, so that other will know how to most effectively apply the consistency regularization to their particular GAN. Finally, we conducted experiments examining *why* consistency regularization improves GAN performance, finding that the answer is not simply that it reduces discriminator over-fitting. Consistency regularization enforces robustness to data perturbation and gives a better training signal for learning the generator. We hope that the proposed consistency regularization will become a standard element in the GAN 'toolkit', and that it will make things easier and simpler for researchers and practitioners.

References

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein GAN. *arXiv:1701.07875*, 2017.

Mikhail Belkin, Partha Niyogi, and Vikas Sindhwani. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *Journal of machine learning research*, 7 (Nov):2399–2434, 2006.

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

Olivier Chapelle, Bernhard Scholkopf, and Alexander Zien. Semi-supervised learning (chapelle, o. et al., eds.; 2006)[book reviews]. *IEEE Transactions on Neural Networks*, 20(3):542–542, 2009.

Emily L. Denton, Soumith Chintala, Arthur Szlam, and Rob Fergus. Deep generative image models using a laplacian pyramid of adversarial networks. In *NIPS*, 2015.

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

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. Generative adversarial nets. In *NIPS*, 2014.

Ishaan Gulrajani, Faruk Ahmed, Martín Arjovsky, Vincent Dumoulin, and Aaron C. Courville. Improved training of wasserstein GANs. In *NIPS*, 2017.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

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 *Advances in Neural Information Processing Systems*, pp. 6626–6637, 2017.

Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. In *ICLR*, 2018.

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

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, 2009.

Karol Kurach, Mario Lucic, Xiaohua Zhai, Marcin Michalski, and Sylvain Gelly. The GAN landscape: Losses, architectures, regularization, and normalization. In *NIPS*, 2019a.

Karol Kurach, Mario Lučić, Xiaohua Zhai, Marcin Michalski, and Sylvain Gelly. A large-scale study on regularization and normalization in gans. In *International Conference on Machine Learning*, pp. 3581–3590, 2019b.

Jae Hyun Lim and Jong Chul Ye. Geometric GAN. *arXiv:1705.02894*, 2017.

Takeru Miyato and Masanori Koyama. cGANs with projection discriminator. In *ICLR*, 2018.

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

Augustus Odena. Open questions about generative adversarial networks. *Distill*, 2019. doi: 10.23915/distill.00018. https://distill.pub/2019/gan-open-problems.

Augustus Odena, Christopher Olah, and Jonathon Shlens. Conditional image synthesis with auxiliary classifier gans. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 2642–2651. JMLR. org, 2017.

Avital Oliver, Augustus Odena, Colin A Raffel, Ekin Dogus Cubuk, and Ian Goodfellow. Realistic evaluation of deep semi-supervised learning algorithms. In *Advances in Neural Information Processing Systems*, pp. 3235–3246, 2018.

Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. In *ICLR*, 2016.

Mehdi Sajjadi, Mehran Javanmardi, and Tolga Tasdizen. Regularization with stochastic transformations and perturbations for deep semi-supervised learning. In *Advances in Neural Information Processing Systems*, pp. 1163–1171, 2016.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *Advances in neural information processing systems*, pp. 2234–2242, 2016.

Dustin Tran, Rajesh Ranganath, and David M. Blei. Deep and hierarchical implicit models. *arXiv:1702.08896*, 2017.

Qizhe Xie, Zihang Dai, Eduard Hovy, Minh-Thang Luong, and Quoc V Le. Unsupervised data augmentation for consistency training. *arXiv preprint arXiv:1904.12848*, 2019.

Chengxi Ye, Yezhou Yang, Cornelia Fermuller, and Yiannis Aloimonos. On the importance of consistency in training deep neural networks. *arXiv preprint arXiv:1708.00631*, 2017.

Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In *BMVC*, 2016.

Xiaohua Zhai, Avital Oliver, Alexander Kolesnikov, and Lucas Beyer. SΘmathbf $\{4\}$ $l : Self - supervised semi - supervised learning.$ *arXiv preprint arXiv:1905.03670*, 2019.

Han Zhang, Ian J. Goodfellow, Dimitris N. Metaxas, and Augustus Odena. Self-attention generative adversarial networks. *arXiv:1805.08318*, 2018.