MEASURING GAN TRAINING IN REAL TIME

Anonymous authors
Paper under double-blind review

ABSTRACT

Generative Adversarial Networks (GAN) are popular generative models of images. Although researchers proposed variants of GAN for different applications, evaluating and comparing GANs is still challenging as GANs may have many failure cases such as low visual quality and model collapse. To alleviate this issue, we propose a novel framework to evaluate the *training stability* (S), *visual quality* (Q), and *mode diversity* (D) of GAN simultaneously. SQD requires only a moderate number of samples, allowing real-time monitoring of the training dynamics of GAN. We showcase the utility of the SQD framework on prevalent GANs and discovered that the gradient penalty (Gulrajani et al., 2017) regularization significantly improves the performance of GAN. We also compare the gradient penalty regularization with other regularization methods and reveal that enforcing the 1-Lipschitz condition of the discriminator network stabilizes GAN training.

1 INTRODUCTION

Generative Adversarial Networks (GANs) (Goodfellow et al., 2014) are a class of generative models for creating new data samples that resemble the training data samples. GANs have raised great interest due to their ability to learn the high-dimensional distribution of images. They have been successful applied to high resolution images generation (Karras et al., 2018; Wang et al., 2018; Brock et al., 2018; Donahue & Simonyan, 2019), style transfer (Zhu et al., 2017a;b; Isola et al., 2017), and text to image synthesis (Reed et al., 2016; Zhang et al., 2017; 2018).

However, training GAN is notoriously elusive. One difficulty is that the loss functions of GAN usually oscillate during training and do not indicate the convergence of the neural networks. Another difficulty is that we lack quantitative assessment methods that measure the visual quality and the mode diversity of the generated images in real time. Without effective and efficient evaluation metrics, practitioners have to visually inspect many generated samples per a few training iterations to decide subjectively when to stop and restart the training iterations. Hence, with an increasing number of literature in GANs (Radford et al., 2015; Zhao et al., 2016; Srivastava et al., 2017; Odena et al., 2017; Kodali et al., 2017; Mao et al., 2017; Arjovsky et al., 2017; Gulrajani et al., 2017; Berthelot et al., 2017; Lim & Ye, 2017; Lin et al., 2018; Arora et al., 2018; Lin et al., 2018; Miyato et al., 2018; Wei et al., 2018; Adler & Lunz, 2018; Bińkowski et al., 2018), the flourishing community calls for good, comprehensive evaluation metrics.

Related Work. The popular Inception Score (IS) (Salimans et al., 2016) and Fréchet Inception Distance (FID) score (Heusel et al., 2017) have adopted a pretrained Inception Network (Szegedy et al., 2015) to extract 2048 embedding features of the generated images to quantitatively evaluate GAN. However, as argued by Sajjadi et al. (2018), Inception Score, FID score and many other evaluation metrics (Theis et al., 2015; Che et al., 2016; Salimans et al., 2016; Tolstikhin et al., 2017; Li et al., 2017; Gurumurthy et al., 2017; Heusel et al., 2017; Arora et al., 2018; Khrulkov & Oseledets, 2018; Bińkowski et al., 2018) provide one-dimensional outputs. They cannot distinguish a GAN between low mode diversity (mode collapse) and low visual quality. Later work (Lucic et al., 2018; Shmelkov et al., 2018; Sajjadi et al., 2018) have proposed a two-dimensional measure "precision and recall" to evaluate visual quality and mode diversity of GAN separately. However, due to the computational complexity, the precision and recall measures cannot serve as real-time indicators for the performance of GAN during training.

Our Methods. We propose a new framework that evaluates the *training stability* (S), *visual quality* (Q), and *mode diversity* (D) simultaneously. We call it the SQD framework. For the training stability, we monitor the angular difference of the network parameters instead of the loss curves. As for the visual quality and the mode diversity, we use two different evaluation metrics separately. We tested the proposed SQD framework on popular variants of GAN (Goodfellow et al., 2014; Mao et al., 2017; Lim & Ye, 2017; Tran et al., 2017; Arjovsky et al., 2017) that model the human faces in the CelebA dataset (Liu et al., 2015). Our contributions are as follows.

- The visual quality and mode diversity metrics in the SQD framework require much fewer samples than previous evaluation methods (Salimans et al., 2016; Heusel et al., 2017; Shmelkov et al., 2018; Sajjadi et al., 2018), and thus allow use to monitor the performance of GAN in real time.
- Our SQD framework reveals that a *turning-point* exists among *all* the tested GANs. That is, after some stable training epochs, the visual quality, mode diversity and angular difference start to deteriorate simultaneously and barely improve in future training.
- Using the SQD framework, we show that the gradient penalty (Gulrajani et al., 2017, GP) regularization, which was originally proposed for WGAN (Arjovsky et al., 2017), addresses the turning-points issue of GANs and improves the visual quality and the mode diversity of GANs. Moreover, we find that GP plays the enhancer role even after the turning points of GANs and that its effect is insensitive to different values of the regularization coefficient.
- We extend the assessment analysis to more regularization methods that share the spirit of GP to enforce the Lipschitz condition of the discriminator network (Kodali et al., 2017; Miyato et al., 2018) and observe similar improvement patterns of GANs.

2 Background

GANs. A GAN typically consists of two neural networks. One is a generator network $G_\theta : z \mapsto x$ with parameter θ that translates a random noise vector z drawn from some distribution $\mathbb{P}_z$ to an image x. The other is a discriminator network $D_\eta(x)$ with parameter η that assigns each image x a probability score of being real. Given a training set of real images $x \sim \mathbb{P}_{\text{data}}$, this generator network G_θ tries to create seemingly realistic image. The discriminator network D_η tries to distinguish real images $x \sim \mathbb{P}_{\text{data}}$ from generated images $G_\theta(z)$ with $z \sim \mathbb{P}_z$.

Formally, the discriminator D_η minimizes a prescribed loss function $L_D : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ given by

$$\min_\eta L_D(D_\eta(x), D_\eta(G_\theta(z))). \tag{1}$$

On the other hand, the generator G_θ is trained to fool the discriminator D_η by minimizing some loss function $L_G : \mathbb{R} \times \mathbb{R} \mapsto \mathbb{R}$ (which can be different from the opposite of the discriminator loss) given by

$$\min_\theta L_G(D_\eta(x), D_\eta(G_\theta(z))). \tag{2}$$

Variants of GAN with different combinations of L_D and L_G have been proposed (Table 1).

Model	Discriminator Loss L_D	Generator Loss L_G
NSGAN	$\mathop{\mathbb{E}}_{x\sim\mathbb{P}_{\text{data}}}[-\log(\sigma(D(x)))] + \mathop{\mathbb{E}}_{z\sim\mathbb{P}_z}[-\log(1-\sigma(D(G(z))))]$	$-\mathop{\mathbb{E}}_{z\sim\mathbb{P}_z}[\log(\sigma(D(G(z))))]$
LSGAN	$\mathop{\mathbb{E}}_{x\sim\mathbb{P}_{\text{data}}}\left[(D(x)-1)^2\right] + \mathop{\mathbb{E}}_{z\sim\mathbb{P}_z}\left[(D(G(z)))^2\right]$	$-\mathop{\mathbb{E}}_{z\sim\mathbb{P}_z}\left[-(D(G(z))-1)^2\right]$
Hinge	$\mathop{\mathbb{E}}_{x\sim\mathbb{P}_{\text{data}}}[\max(0, 1-D(x))] + \mathop{\mathbb{E}}_{z\sim\mathbb{P}_z}[\max(0, 1+D(G(z)))]$	$-\mathop{\mathbb{E}}_{z\sim\mathbb{P}_z}[D(G(z))]$
WGAN	$\mathop{\mathbb{E}}_{x\sim\mathbb{P}_{\text{data}}}[-D(x)] + \mathop{\mathbb{E}}_{z\sim\mathbb{P}_z}[D(G(z))]$	$-\mathop{\mathbb{E}}_{z\sim\mathbb{P}_z}[D(G(z))]$

Table 1: Discriminator and generator loss functions of non-saturating GAN (Goodfellow et al., 2014, NSGAN), least square GAN (Mao et al., 2017, LSGAN), Hinge Loss GAN (Lim & Ye, 2017; Tran et al., 2017), Wasserstein GAN (Arjovsky et al., 2017, WGAN). For NSGAN, $\sigma(t) = (1+e^{-t})^{-1} \in (0, 1)$ is the sigmoid function. A similar table can be found in Lucic et al. (2018).

Gradient Penalty (GP). Proposed by Gulrajani et al. (2017), GP encourages $\nabla_\eta D_\eta(\hat{x})$ to be of a unit length by regularizing the discriminator loss L_D of WGAN (Arjovsky et al., 2017, see the last row of Table 1) with a penalty term

$$\mathbb{E}_{\hat{x}\sim\mathbb{P}_{\hat{x}}}\left[\left(\left\|\nabla_\eta D_\eta(\hat{x})\right\|_2 - 1\right)^2\right], \tag{3}$$

where each $\hat{x}$ is a convex combination of a real image x and a generated image $G(z)$ with the combination coefficient following a uniform distribution between $[0, 1]$

$$\hat{x} = \alpha x + (1-\alpha)G(z), \quad \alpha \sim \text{Uniform}[0, 1].$$

In this paper, we would showcase the utility of our SQD framework by investigating whether and to what extent GP improves the performances of four variants of GANs listed in Table 1.

3 Methods

We describe how the SQD framework evaluates the training stability, visual quality, and mode diversity of GANs. We showcase our methodology on GANs that model human faces in the CelebA dataset (Liu et al., 2015). The CelebA dataset contains 202 599 face images of 10 177 celebrities.

Training Stability. Instead of using the unstable loss curves, we track the angular difference of the networks' weight parameters. Let $\boldsymbol{W}_t^{(G)}$ and $\boldsymbol{W}_t^{(D)}$ be the weight matrices of the generator and discriminator networks at iteration t, respectively. Define the angle between two weight matrices $\boldsymbol{W}_t$, $\boldsymbol{W}_{t+1}$ as

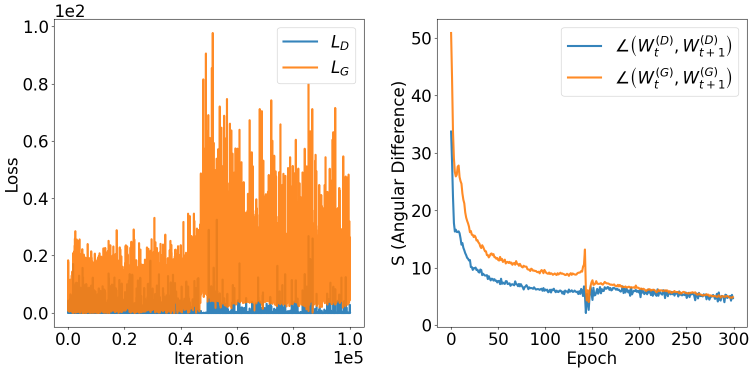
$$\angle\left(\boldsymbol{W}_t, \boldsymbol{W}_{t+1}\right) = \arccos\left(\frac{\text{vec}(\boldsymbol{W}_t)^\top \text{vec}(\boldsymbol{W}_{t+1})}{\|\text{vec}(\boldsymbol{W}_t)\|_2 \, \|\text{vec}(\boldsymbol{W}_{t+1})\|_2}\right). \tag{4}$$

We monitor the angles

$$\text{S} := \left(\angle\left(\boldsymbol{W}_t^{(G)}, \boldsymbol{W}_{t+1}^{(G)}\right), \angle\left(\boldsymbol{W}_t^{(D)}, \boldsymbol{W}_{t+1}^{(D)}\right)\right) \tag{5}$$

at each iteration t and expect their convergences towards 0 as $t \to \infty$, which indicates training stability.

The reasons for examining the angular difference between weight parameters instead of the loss curves are as follows. The loss curves of both D_η and G_θ oscillate dramatically and thus cannot diagnose convergence during training. Recent progress on understanding the implicit bias of deep neural networks (Soudry et al., 2018; Gunasekar et al., 2018; Moroshko et al., 2020) has suggested that the directions of the network parameters converge with training iterations. Figure 1 plots the losses and the angular difference between weight parameters of an NSGAN that is trained on the CelebA dataset. The loss curves oscillate throughout the whole training process, but the angular difference converges.



(a) Losses (b) Angular Differences

Figure 1: Loss curves (a) and angular difference curves (b) during 300 training epochs of a NSGAN on the CelebA dataset. The clear jumps of the angular difference curves in (b) correspond to a qualitative transition of the NSGAN network as we will discuss later.

Visual Quality. We use the MTCNN face detection network (Zhang et al., 2016) to judge the visual quality of generated images. The MTCNN network outputs a visual quality score $\mathrm{Q}(x) \in [0, 1]$ for each image x, indicating the probability that this image is a human face. Suppose that a variant of GAN generates a set of images $\hat{\mathcal{X}}$. We define the average probability score assigned by MTCNN to each generated image as the visual quality score of the GAN.

$$\mathrm{Q} := |\hat{\mathcal{X}}|^{-1} \sum_{\hat{x} \in \hat{\mathcal{X}}} \mathrm{Q}(\hat{x}) \tag{6}$$

The advantages of the visual quality score in Equation 6 over the popular Inception Score and FID score are twofold. First, both Inception Score and FID scores use a feature embedding network designed for general visual perception but our visual quality scores is given by a perception network trained specifically for human face recognition. Second, both Inception Score and FID scores require a large number of generated images to evaluate because they need to quantify the divergence between two distributions in the 2048-dimensional embedding space of the Inception Network. The variance of our visual quality score is at most $0.25 \times |\hat{\mathcal{X}}|^{-1}$. A small sample of 1000 generated images leads to an accurate estimate of the visual quality score with a standard deviation ≤ 0.016.

Mode Diversity. Mode collapse is a major challenge in GAN: only a few modes (subpopulations) in the target distribution $\mathbb{P}_{\text{data}}$ are learned by GAN and overrepresented in the generated images, while other modes are underrepresented or even dropped. Effective evaluation metrics for mode collapse had been missing in the literature since the birth of GAN. Recently, Sajjadi et al. (2018) generalized the concept of "precision and recall" from the context of binary classification to that of generative modeling, and measured the degree of mode dropping by the portion of the target distribution not covered by the learned distribution. This metric considers whether various modes of the target distribution are present or absent in the learned distribution.

We step further and evaluate whether various modes of the target distribution are underrepresented or overrepresented in the learned distribution. Recall that the CelebA dataset consists of 202 599 face images of 10 177 celebrities. We view the identity of each celebrity as a mode (subpopulation) in the training data, and use the ArcFace face verification network (Deng et al., 2019) to find the most similar identity for each generated face image. The ArcFace network is designed to verify whether two face images belong to the same person. Let $\mathcal{X}_j$ be the subset of face images belonging to the identity $j \in \{1, \ldots, m\}$ with $m = 10\,177$ being the number of modes, and let f_j be the mean feature vector of its images in the embedding space of the ArcFace network.

$$f_j = |\mathcal{X}_j|^{-1} \sum_{x \in \mathcal{X}_j} f(x), \quad j = 1, \ldots, m, \tag{7}$$

where $f : \mathcal{X} \to \mathbb{R}^d$ is the ArcFace embedding network. For each generated image $\hat{x}$, we assign it to its most similar identity

$$j(\hat{x}) = \arg\min_j \left\| f(\hat{x}) - f_j \right\|_2 . \tag{8}$$

The entropy of $j(\hat{x})$ reflects the degree of mode underrepresentation and overrepresentation. We define it as the mode diversity of GAN

$$\mathrm{D} := \mathrm{Entropy}(j(\hat{x})). \tag{9}$$

Moreover, since this mode diversity metric requires only $n \asymp m/\log m$ samples, it can indicate mode diversity or mode collapse in real time during the GAN training. Indeed, theoretical results on the entropy estimation for a discrete distribution suggest that the sample size $n \asymp m/\log m$ is necessary and sufficient for consistently estimating the entropy at the optimal rate (Valiant & Valiant, 2011; Jiao et al., 2015; Wu & Yang, 2016). For the CelebA dataset, $m = 10\,177$ and $m/\log m \approx 1103$. In our experiments, we adopted the James-Stein entropy estimation method (Hausser & Strimmer, 2009) and found that a sample size $n = 2000$ worked well.

For n generated images $\{\hat{x}_i\}_{i=1}^n$, let $\mathcal{J} = \{j_i = j(\hat{x}_i)\}_{i=1}^n$ be the set of the indices of their most similar identities. Algorithm 1 describes the James-Stein method.

4 Experiments

We considered four variants of GAN — NSGAN, Hinge-loss GAN, LSGAN and WGAN (Table 1) — with and without the GP regularization, respectively, on the discriminator. The loss function of the

Algorithm 1 James-Stein Entropy Estimation

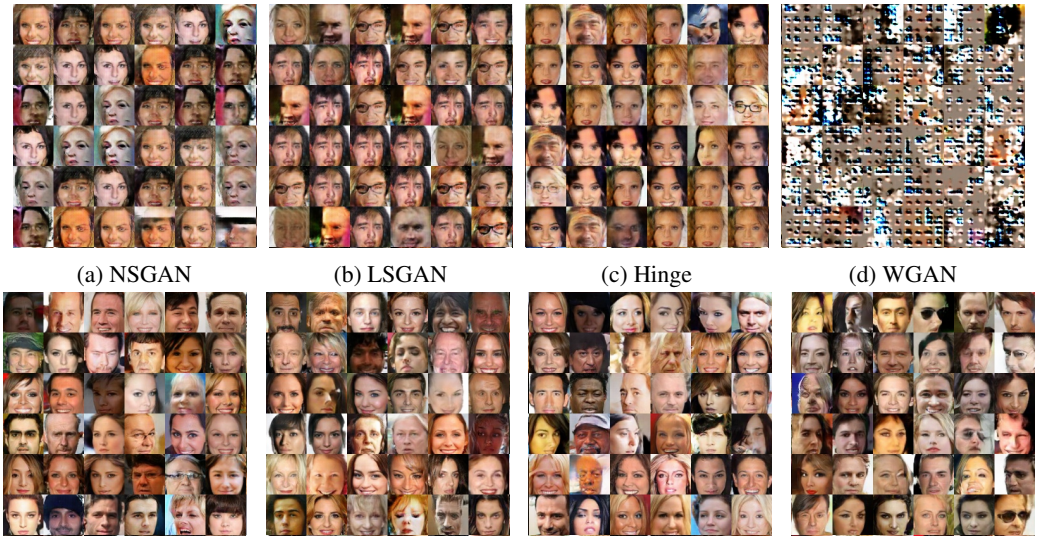
1: **Input** $\mathcal{J} = \{j_1, j_2, \ldots, j_n\}$.
2: $p_{n,j} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{j_i=j\}},\ \forall j = 1, ..., m.$
3: $\lambda = \left(1 - \sum_{j=1}^{m} p_{n,j}^2\right) \Big[(n-1)\sum_{j=1}^{m}(m^{-1} - p_{n,j})^2\Big].$
4: $\hat{p}_{n,j} = \lambda m^{-1} + (1-\lambda)p_{n,j}.$
5: **Output** $-\sum_{j=1}^{m} \hat{p}_{n,j} \log \hat{p}_{n,j}.$

discriminator regularized by GP was

$$L_{D,\lambda} = L_D + \lambda \mathbb{E}_{\hat{x}\sim\mathbb{P}_{\hat{x}}} \left[\left(\left\|\nabla_\eta D_\eta(\hat{x})\right\|_2 - 1\right)^2\right]. \tag{10}$$

We set the regularization coefficient λ to 10 unless specified, and adopted the ADAM optimizer (Kingma & Ba, 2014) to train all GANs in comparison.

Figure 2 visualizes 100 generated images for each of eight GANs in comparison after 300 training epochs. Mode collapse arose in NSGAN, Hinge-loss GAN, and LSGAN. Images generated by WGAN were of low visual quality. For each of these four GANs, adding the GP regularization to its discriminator improved both visual quality and mode diversity.



(a) NSGAN (b) LSGAN (c) Hinge (d) WGAN

(e) NSGAN-GP (f) LSGAN-GP (g) Hinge-GP (h) WGAN-GP

Figure 2: Generated images of GANs after 300 training epochs.

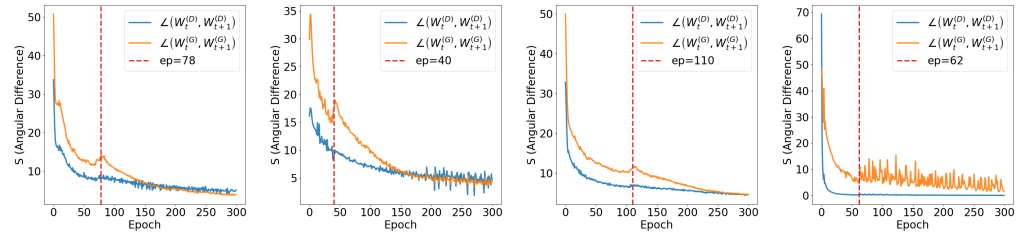
Interestingly, with the SQD framework, we observed that the training dynamic of each GAN without the GP regularization exhibited a turning point when the network became unstable and severe mode collapse occurred (Figures 3-5). In contrast, each GAN with the GP regularization did not suffer from such issue.

Training Stability. As we have discussed in Section 3, the angular difference of both the generator network G and the discriminator network D should converge to 0 smoothly if GAN improves incrementally and stably. Figure 3 shows that GANs without the GP regularization were highly unstable during training. Adding the GP regularization to the discriminator of GAN stablized training.

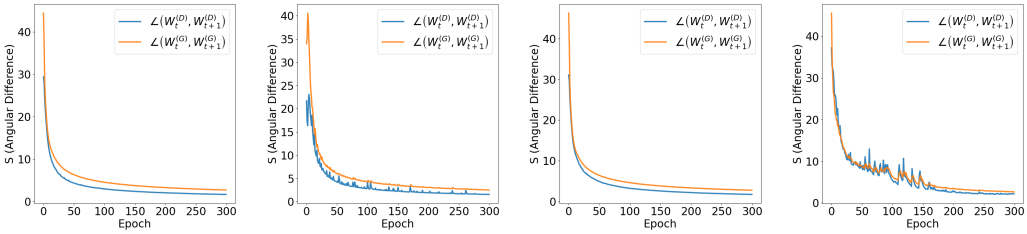
- The weight parameter of the NSGAN's generator network dramatically changed direction at epoch 78. Similar changes happened at epoch 40 for LSGAN, epoch 110 for Hinge-loss GAN, and epoch 62 for WGAN. The weight parameters of both the generator and discriminator networks smoothly converged.
- The training dynamic of WGAN was overall less stable than the other three GANs. Without GP, the weight parameter of its generator frequently changed directions even after the turning

point and did not converge. With GP, the weight parameters suddenly changed directions only occasionally.

- The training instability of GANs with no GP caused dramatic changes of visual quality (Figure 4) and severe mode collapses (Figure 5) at turning points.

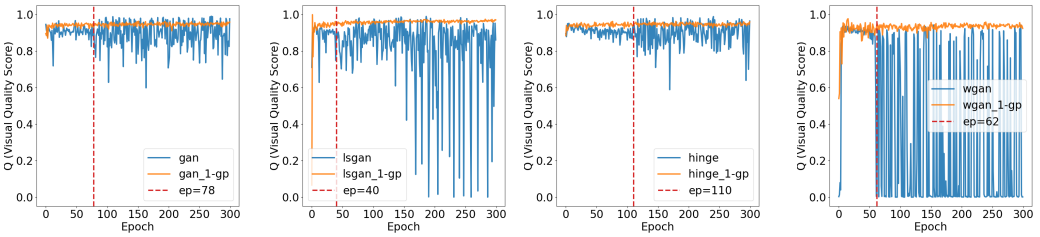


(a) NSGAN (b) LSGAN (c) Hinge (d) WGAN



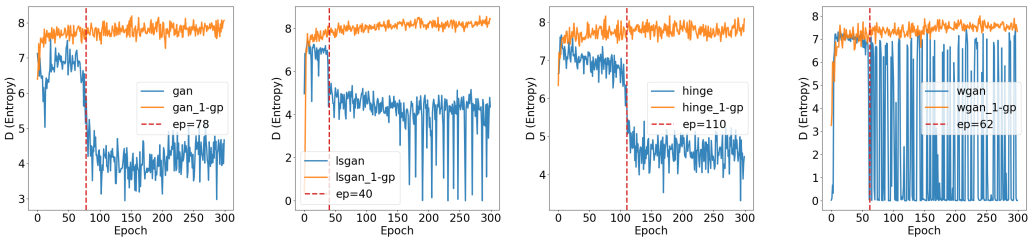
(e) NSGAN-GP (f) LSGAN-GP (g) Hinge-GP (h) WGAN-GP

Figure 3: Angular differences of different GANs



(a) NSGAN (b) LSGAN (c) Hinge (d) WGAN

Figure 4: Visual quality of different GANs w/ and w/o GP regularization.



(a) NSGAN (b) LSGAN (c) Hinge (d) WGAN

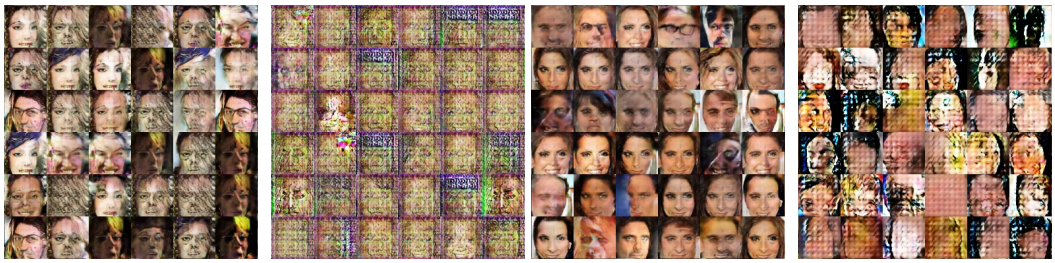
Figure 5: Mode diversity of different GANs w/ and w/o GP regularization.

Visual Quality. Figure 4 shows that with no GP regularization, the visual quality scores of GANs increased stably in the early training stage but suddenly started to oscillate at epoch 78 of NSGAN, epoch 40 of LSGAN, epoch 110 of Hinge-loss GAN, and epoch 62 of WGAN. This signifies the low or poor visual quality of the generated images. In contrast, GANs with GP evolved smoothly throughout all the 300 training epochs and gave high visual quality scores consistently. The visual quality scores of WGAN frequently hit zero after the turning point, indicating that WGAN created

images of extremely low quality. This aligns well with the human visual judgement on generated images of WGAN (Figure 2(d)).

Mode Diversity. Figure 5 shows that with no GP regularization, the mode diversity scores of NSGAN, LSGAN and Hinge-loss GAN jumped at their turning points, indicating severe mode collapse (Figure 6). The large swings of the mode diversity score of WGAN between 0 and 7 after the turning point epoch 62 reconfirms the training instability of WGAN. GANs with GP regularization enjoyed relatively good mode diversity. Their mode diversity scores higher than 7.50 are reasonably close to its theoretically maximum value $\ln(10\,177) = 9.23$ for the CelebA dataset.

Figure 6 shows sample generated images of each GAN at its turning point. The appendix shows more generated images at different epochs to illustrate the evolution of GANs.



(a) NSGAN at epoch 74 (b) LSGAN at epoch 40 (c) Hinge at epoch 110 (d) WGAN at epoch 62

Figure 6: Generated images of GANs at their turning points

Our evaluation shows that the SQD framework is an effective diagnostic toolbox to monitor the training process of GAN in real time.

5 Analysis of the Gradient Penalty Regularization

We average the visual quality and mode diversity scores of GANs during epochs 250–300 and conclude that GP regularization significantly improves the performance of GANs.

Model	Q w/o GP	Q w/ GP	D w/o GP	D w/ GP
NSGAN	0.893	**0.949**	4.289	**7.848**
LSGAN	0.899	**0.946**	4.583	**7.794**
Hinge	0.787	**0.964**	3.963	**8.221**
WGAN	0.243	**0.937**	1.885	**7.556**

Table 2: The average visual quality (Q) and mode diversity (D) of the last 50 epochs.

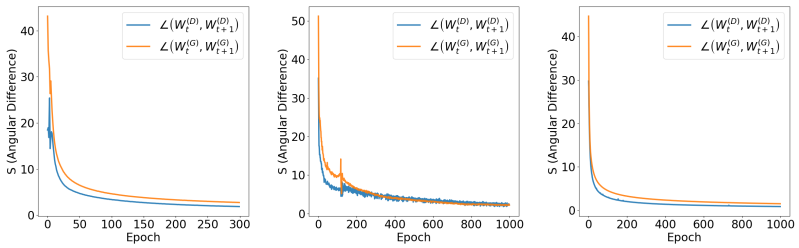
We conduct more experiments to investigate how GP improves the training stability of GAN. We first increased the regularization coefficient λ from 10 to 100 and found that, despite the minor fluctuation in the earlier training stage, GP with a larger weight $(\lambda = 100)$ still stabilized NSGAN training (Figure 7(a)). This finding confirms the claim of Petzka et al. (2018) that the advantages of GP holds for a wide range of λ.

Next, we monitored more training iterations of NSGAN up to 1000 epochs (Figure 7 (b–c)), and found that GP kept stabilizing the training dynamic of NSGAN.

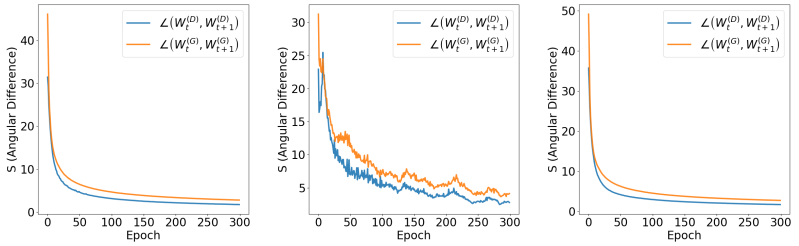
Finally, we extended the assessment analysis to three other regularization methods that have similar forms to GP, namely, Dragan, Lipschitz and 0-GP (Table 3). Interestingly, we found that Dragan and Lipschitz methods stabilized GAN training in ways similar to GP, but 0-GP destabilized the training (Figure 7(d–f)). A possible explanation is that they share the same spirit to regularize the Lipschitz condition of networks. Miyato et al. (2018) justified the effect of the Lipschitz condition on the training stability of GAN.

Regularization	Expression	$\mathbb{P}_{\hat{x}}$
GP	$\lambda \underset{\hat{x}\sim\mathbb{P}_{\hat{x}}}{\mathbb{E}} \left[(\|\nabla_\eta D_\eta(\hat{x})\|_2 - 1)^2\right]$	$\hat{x} = \alpha x + (1-\alpha)G(z), \alpha \in [0,1]$
Dragan	$\lambda \underset{\hat{x}\sim\mathbb{P}_{\hat{x}}}{\mathbb{E}} \left[(\|\nabla_\eta D_\eta(\hat{x})\|_2 - 1)^2\right]$	$\hat{x} = x + \delta, \delta \sim N(0, c\boldsymbol{I})$
0-GP	$\lambda \underset{\hat{x}\sim\mathbb{P}_{\hat{x}}}{\mathbb{E}} \left[\|\nabla_\eta D_\eta(\hat{x})\|_2^2\right]$	$\hat{x} = \alpha x + (1-\alpha)G(z), \alpha \in [0,1]$
Lipschitz	$\lambda \underset{\hat{x}\sim\mathbb{P}_{\hat{x}}}{\mathbb{E}} \left[(\max\{0, \|\nabla_\eta D_\eta(\hat{x})\|_2 - 1\})^2\right]$	$\hat{x} = \alpha x + (1-\alpha)G(z), \alpha \in [0,1]$

Table 3: Different Regularizations of GAN: GP (Gulrajani et al., 2017), Dragan (Kodali et al., 2017), 0-GP (Thanh-Tung et al., 2019), Lipschitz (Petzka et al., 2018). $x \sim \mathbb{P}_x$ is the real image, $G(z)$ is the generated image.



(a) GP ($\lambda = 100$) (b) NSGAN 1000 epochs (c) GP 1000 epochs



(d) Dragan (e) 0-GP (f) Lipschitz

Figure 7: Angular difference of NSGAN with different regularizations: Dragan (Kodali et al., 2017), 0-centered gradient penalty (0-GP) (Thanh-Tung et al., 2019), Lipschitz constraint and gradient penalty (Gulrajani et al., 2017). The weight of penalty terms are set to $\lambda = 10$ unless specified.

6 Conclusion and Future Work

We propose the SQD framework to monitor the training stability, visual quality, and mode diversity of GAN training in real time. Our framework uses the angular difference of the network weight matrices to signify training stability. We used a detection network and a recognition network to evaluate the visual quality and mode diversity of the generated images separately. We tested the SQD framework on various GANs for modeling human faces.

In addition, our experiments show that GP (Gulrajani et al., 2017) regularization improves the performance of GANs in training stability, visual quality, and mode diversity. We further investigated regularization methods that have similar forms and found that Dragan and Lipschitz methods stabilize GAN training as well.

As we only conducted experiments on the CelebA dataset, we would like to extent our methods to evaluate GANs trained on other datasets. Besides the discriminator regularizations mentioned in Table 3, many other regularization techniques (Che et al., 2016; Brock et al., 2016; Roth et al., 2017; Miyato et al., 2018; Odena et al., 2018; Chu et al., 2020) have demonstrated the potential in improving GAN training. We would also like to investigate the theoretical justification for these regularizations.

REFERENCES

Jonas Adler and Sebastian Lunz. Banach wasserstein GAN. In *Advances in Neural Information Processing Systems*, pp. 6754–6763, 2018.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International conference on machine learning*, pp. 214–223, 2017.

Sanjeev Arora, Andrej Risteski, and Yi Zhang. Do GANs learn the distribution? some theory and empirics. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=BJehNfW0-.

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

Mikołaj Bińkowski, Dougal J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans. *arXiv preprint arXiv:1801.01401*, 2018.

Andrew Brock, Theodore Lim, James M Ritchie, and Nick Weston. Neural photo editing with introspective adversarial networks. *arXiv preprint arXiv:1609.07093*, 2016.

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

Tong Che, Yanran Li, Athul Paul Jacob, Yoshua Bengio, and Wenjie Li. Mode regularized generative adversarial networks. *arXiv preprint arXiv:1612.02136*, 2016.

Casey Chu, Kentaro Minami, and Kenji Fukumizu. Smoothness and stability in GANs. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=HJeOekHKwr.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4690–4699, 2019.

Jeff Donahue and Karen Simonyan. Large scale adversarial representation learning. In *Advances in Neural Information Processing Systems*, pp. 10541–10551, 2019.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pp. 2672–2680, 2014.

Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein GANs. In *Advances in neural information processing systems*, pp. 5767–5777, 2017.

Suriya Gunasekar, Jason Lee, Daniel Soudry, and Nathan Srebro. Characterizing implicit bias in terms of optimization geometry. *arXiv preprint arXiv:1802.08246*, 2018.

Swaminathan Gurumurthy, Ravi Kiran Sarvadevabhatla, and R Venkatesh Babu. Deligan: Generative adversarial networks for diverse and limited data. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 166–174, 2017.

Jean Hausser and Korbinian Strimmer. Entropy inference and the james-stein estimator, with application to nonlinear gene association networks. *Journal of Machine Learning Research*, 10(7), 2009.

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.

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1125–1134, 2017.

Jiantao Jiao, Kartik Venkat, Yanjun Han, and Tsachy Weissman. Minimax estimation of functionals of discrete distributions. *IEEE Transactions on Information Theory*, 61(5):2835–2885, 2015.

Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=Hk99zCeAb.

Valentin Khrulkov and Ivan Oseledets. Geometry score: A method for comparing generative adversarial networks. *arXiv preprint arXiv:1802.02664*, 2018.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

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

Chun-Liang Li, Wei-Cheng Chang, Yu Cheng, Yiming Yang, and Barnabás Póczos. Mmd gan: Towards deeper understanding of moment matching network. In *Advances in Neural Information Processing Systems*, pp. 2203–2213, 2017.

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

Zinan Lin, Ashish Khetan, Giulia Fanti, and Sewoong Oh. Pacgan: The power of two samples in generative adversarial networks. In *Advances in neural information processing systems*, pp. 1498–1507, 2018.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

Mario Lucic, Karol Kurach, Marcin Michalski, Sylvain Gelly, and Olivier Bousquet. Are GANs created equal? a large-scale study. In *Advances in neural information processing systems*, pp. 700–709, 2018.

Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2794–2802, 2017.

Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=B1QRgziT-.

Edward Moroshko, Suriya Gunasekar, Blake Woodworth, Jason D Lee, Nathan Srebro, and Daniel Soudry. Implicit bias in deep linear classification: Initialization scale vs training accuracy. *arXiv preprint arXiv:2007.06738*, 2020.

Augustus Odena, Christopher Olah, and Jonathon Shlens. Conditional image synthesis with auxiliary classifier GANs. In *International conference on machine learning*, pp. 2642–2651, 2017.

Augustus Odena, Jacob Buckman, Catherine Olsson, Tom B Brown, Christopher Olah, Colin Raffel, and Ian Goodfellow. Is generator conditioning causally related to gan performance? *arXiv preprint arXiv:1802.08768*, 2018.

Henning Petzka, Asja Fischer, and Denis Lukovnikov. On the regularization of wasserstein GANs. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=B1hYRMbCW.

Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. *arXiv preprint arXiv:1605.05396*, 2016.

Kevin Roth, Aurelien Lucchi, Sebastian Nowozin, and Thomas Hofmann. Stabilizing training of generative adversarial networks through regularization. In *Advances in neural information processing systems*, pp. 2018–2028, 2017.

Mehdi SM Sajjadi, Olivier Bachem, Mario Lucic, Olivier Bousquet, and Sylvain Gelly. Assessing generative models via precision and recall. In *Advances in Neural Information Processing Systems*, pp. 5228–5237, 2018.

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.

Konstantin Shmelkov, Cordelia Schmid, and Karteek Alahari. How good is my GAN? In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 213–229, 2018.

Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data. *The Journal of Machine Learning Research*, 19(1): 2822–2878, 2018.

Akash Srivastava, Lazar Valkov, Chris Russell, Michael U Gutmann, and Charles Sutton. VeeGAN: Reducing mode collapse in GANs using implicit variational learning. In *Advances in Neural Information Processing Systems*, pp. 3308–3318, 2017.

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1–9, 2015.

Hoang Thanh-Tung, Truyen Tran, and Svetha Venkatesh. Improving generalization and stability of generative adversarial networks. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=ByxPYjC5KQ.

Lucas Theis, Aäron van den Oord, and Matthias Bethge. A note on the evaluation of generative models. *arXiv preprint arXiv:1511.01844*, 2015.

Ilya O Tolstikhin, Sylvain Gelly, Olivier Bousquet, Carl-Johann Simon-Gabriel, and Bernhard Schölkopf. Adagan: Boosting generative models. In *Advances in Neural Information Processing Systems*, pp. 5424–5433, 2017.

Dustin Tran, Rajesh Ranganath, and David Blei. Hierarchical implicit models and likelihood-free variational inference. In *Advances in Neural Information Processing Systems*, pp. 5523–5533, 2017.

Gregory Valiant and Paul Valiant. Estimating the unseen: an n/log (n)-sample estimator for entropy and support size, shown optimal via new clts. In *Proceedings of the forty-third annual ACM symposium on Theory of computing*, pp. 685–694, 2011.

Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional GANs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 8798–8807, 2018.

Xiang Wei, Boqing Gong, Zixia Liu, Wei Lu, and Liqiang Wang. Improving the improved training of wasserstein GANs: A consistency term and its dual effect. *arXiv preprint arXiv:1803.01541*, 2018.

Yihong Wu and Pengkun Yang. Minimax rates of entropy estimation on large alphabets via best polynomial approximation. *IEEE Transactions on Information Theory*, 62(6):3702–3720, 2016.

Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. StackGAN: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pp. 5907–5915, 2017.

Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, 2016.

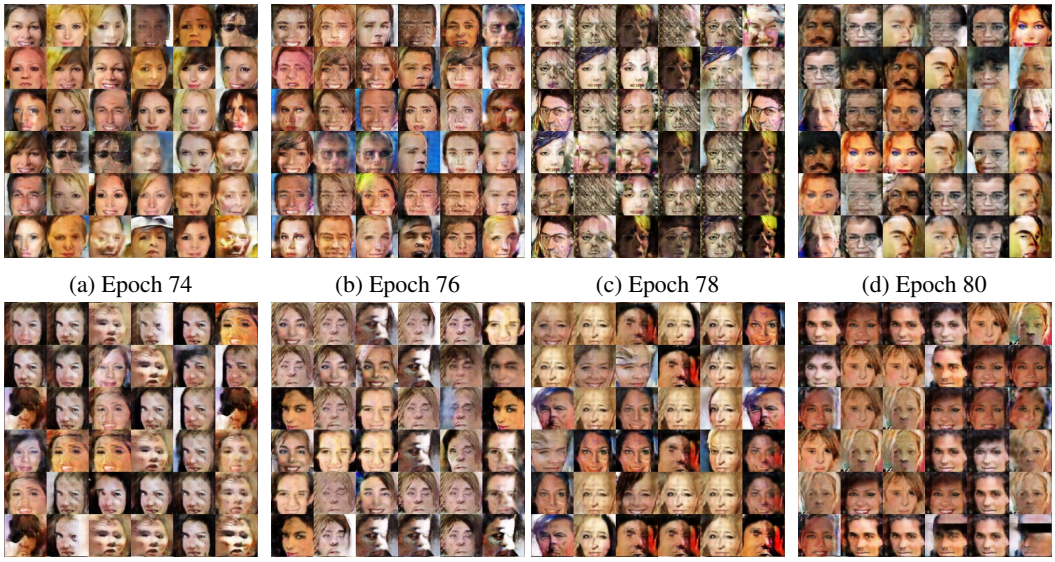
Zizhao Zhang, Yuanpu Xie, and Lin Yang. Photographic text-to-image synthesis with a hierarchically-nested adversarial network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6199–6208, 2018.

Junbo Zhao, Michael Mathieu, and Yann LeCun. Energy-based generative adversarial network. *arXiv preprint arXiv:1609.03126*, 2016.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pp. 2223–2232, 2017a.

Jun-Yan Zhu, Richard Zhang, Deepak Pathak, Trevor Darrell, Alexei A Efros, Oliver Wang, and Eli Shechtman. Toward multimodal image-to-image translation. In *Advances in neural information processing systems*, pp. 465–476, 2017b.

A EVOLUTION OF DIFFERENT GAN MODELS DURING TRAINING



(a) Epoch 74 (b) Epoch 76 (c) Epoch 78 (d) Epoch 80

(e) Epoch 82 (f) Epoch 84 (g) Epoch 86 (h) Epoch 300

Figure 8: Generated samples of NSGAN in different epochs

(a) Epoch 35 (b) Epoch 70 (c) Epoch 105 (d) Epoch 140

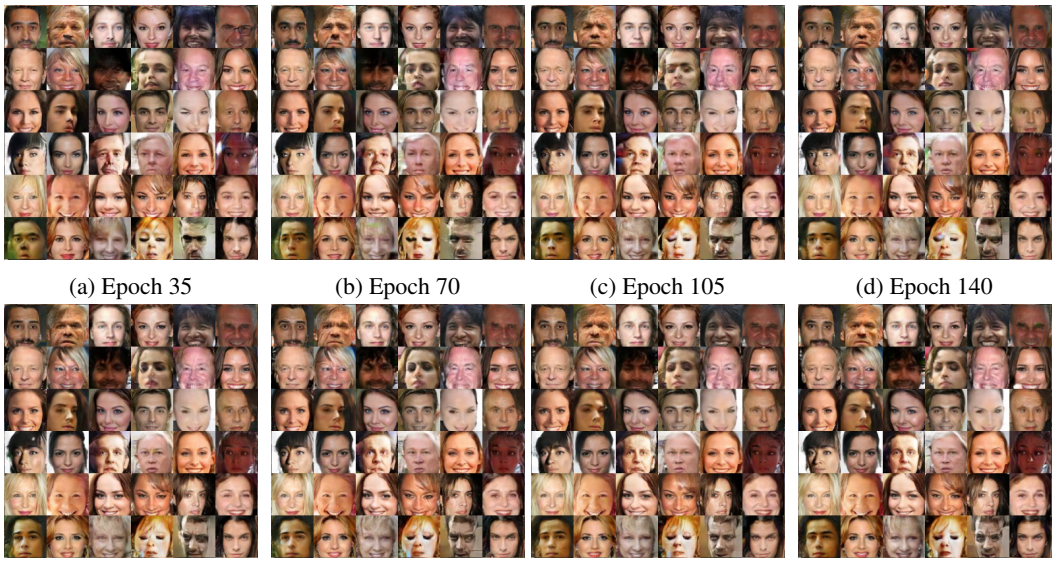
(e) Epoch 175 (f) Epoch 210 (g) Epoch 245 (h) Epoch 280

Figure 9: Generated samples of NSGAN-GP in different epochs

(a) Epoch 36 (b) Epoch 38 (c) Epoch 40 (d) Epoch 42

(e) Epoch 44 (f) Epoch 46 (g) Epoch 48 (h) Epoch 300

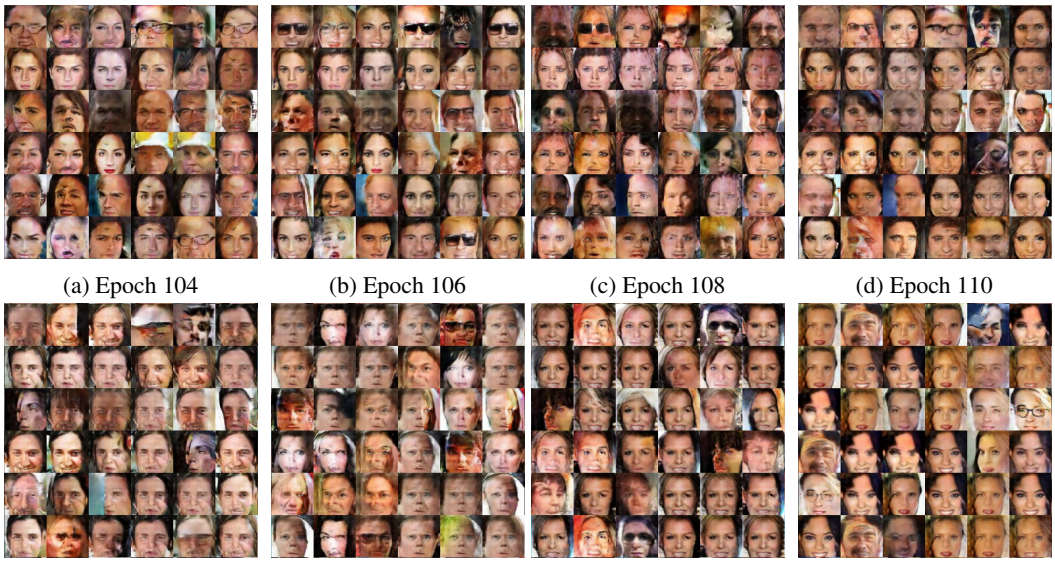
Figure 10: Generated samples of LSGAN in different epochs



(a) Epoch 35 (b) Epoch 70 (c) Epoch 105 (d) Epoch 140

(e) Epoch 175 (f) Epoch 210 (g) Epoch 245 (h) Epoch 280

Figure 11: Generated samples of LSGAN-GP in different epochs



(a) Epoch 104 (b) Epoch 106 (c) Epoch 108 (d) Epoch 110

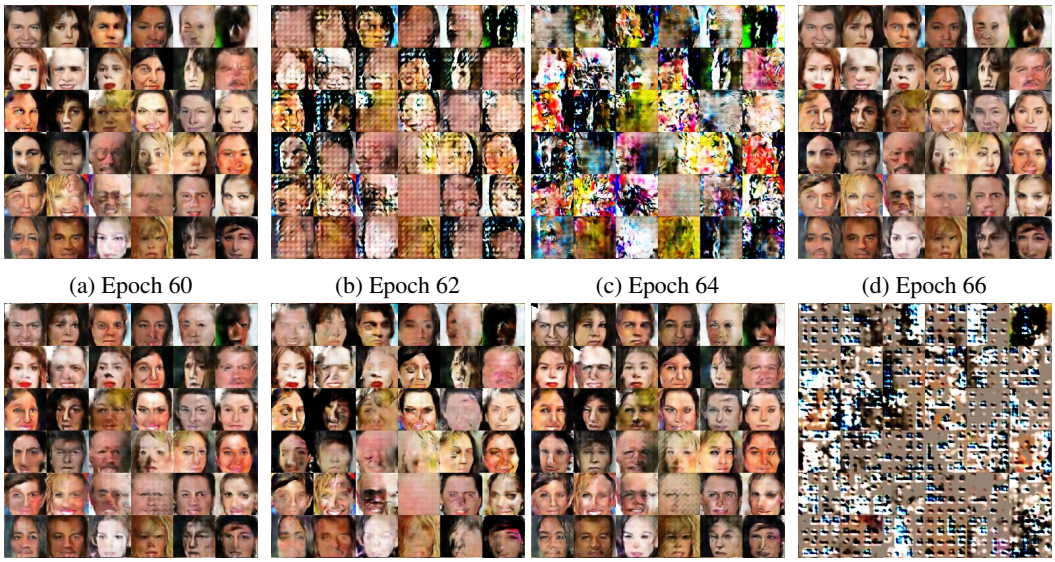
(e) Epoch 112 (f) Epoch 114 (g) Epoch 116 (h) Epoch 300

Figure 12: Generated samples of Hinge Loss in different epochs

(a) Epoch 35 (b) Epoch 70 (c) Epoch 105 (d) Epoch 140

(e) Epoch 175 (f) Epoch 210 (g) Epoch 245 (h) Epoch 280

Figure 13: Generated samples of Hinge-GP in different epochs



(a) Epoch 60 (b) Epoch 62 (c) Epoch 64 (d) Epoch 66

(e) Epoch 68 (f) Epoch 70 (g) Epoch 72 (h) Epoch 300

Figure 14: Generated samples of WGAN in different epochs

(a) Epoch 35 (b) Epoch 70 (c) Epoch 105 (d) Epoch 140

(e) Epoch 175 (f) Epoch 210 (g) Epoch 245 (h) Epoch 280

Figure 15: Generated samples of WGAN-GP in different epochs

B First Order Necessity Condition of $L_{D,\lambda}$

Here we provide the first order condition of the critical point of the GAN objective function $L_{D,\lambda}$ with the gradient penalty term $\lambda \mathbb{E}_{x\sim\mathbb{P}_{\hat{x}}}\left(\|\nabla_\eta D_\eta(\tilde{x})\|_2 - 1\right)^2$. We abuse the sampling issue of $G_\theta, D_\eta, \nabla_\eta D_\eta$ and write the GAN-GP discriminator loss as

$$\min_\eta L_{D,\lambda} = L_D + \lambda \left(\|\nabla_\eta D_\eta\|_2 - 1\right)^2. \tag{11}$$

With the gradient penalty, the first order critical points of generator G_θ remains the same:

$$\frac{\partial}{\partial \eta} L_{D,\lambda} = 0. \tag{12}$$

The first order critical points of the discriminator D_η satisfy:

$$\begin{aligned} & \frac{\partial}{\partial \eta}\left[L_D(D_\eta, D_\eta(G_\theta)) + \lambda \left(\|\nabla_\eta D_\eta\|_2 - 1\right)^2\right] = 0 \\ \Longrightarrow\ & \frac{dL_D}{dD}\nabla_\eta D_\eta + 2\lambda \nabla_\eta^2 D_\eta \nabla_\eta D_\eta \left(1 - \frac{1}{\|\nabla_\eta D_\eta\|_2}\right) = 0 \\ \Longrightarrow\ & \left[\frac{dL_{D,\lambda}}{dD}\boldsymbol{I} + 2\lambda\left(1 - \frac{1}{\|\nabla_\eta D_\eta\|_2}\right)\nabla_\eta^2 D_\eta\right]\nabla_\eta D_\eta = 0 \end{aligned} \tag{13}$$

where $\nabla_\eta^2 D_\eta$ is the Hessian matrix of D_η. This critical condition equation 13 suggests $\nabla_\eta D_\eta$ should be in the null space of matrix

$$\frac{dL_{D,\lambda}}{dD}\boldsymbol{I} + 2\lambda\left(1 - \frac{1}{\|\nabla_\eta D_\eta\|_2}\right)\nabla_\eta^2 D_\eta. \tag{14}$$