Cost Ensemble with Gradient Selecting for GANs

Minghui Liu, **Jiali Deng**, **Meiyi Yang**, **Xuan Cheng**, **Nianbo Liu**, **Ming Liu** and **Xiaomin Wang***
University of Electronic Science and Technology of China
minghuiliu@std.uestc.edu.cn, dengjiali@std.uestc.edu.cn, meiyiyang@std.uestc.edu.cn, cs_xuancheng@std.uestc.edu.cn, liunb@uestc.edu.cn, csmliu@uestc.edu.cn, xmwang@uestc.edu.cn

Abstract

Generative Adversarial Networks(GANs) are powerful generative models on numerous tasks and datasets but are also known for their training instability and mode collapse. The latter is because the optimal transportation map is discontinuous, but DNNs can only approximate continuous ones. One way to solve the problem is to introduce multiple discriminators or generators. However, their impacts are limited because the cost function of each component is the same. That is, they are homogeneous. In contrast, multiple discriminators with different cost functions can yield various gradients for the generator, which indicates we can use them to search for more transportation maps in the latent space. Inspired by this, we have proposed a framework to combat the mode collapse problem, containing multiple discriminators with different cost functions, named CES-GAN. Unfortunately, it may also lead to the generator being hard to train because the performance between discriminators is unbalanced, according to the Cannikin Law. Thus, a gradient selecting mechanism is also proposed to pick up proper gradients. We provide mathematical statements to prove our assumptions and conduct extensive experiments to verify the performance. The results show that CES-GAN is lightweight and more effective for fighting against the mode collapse problem than similar works.

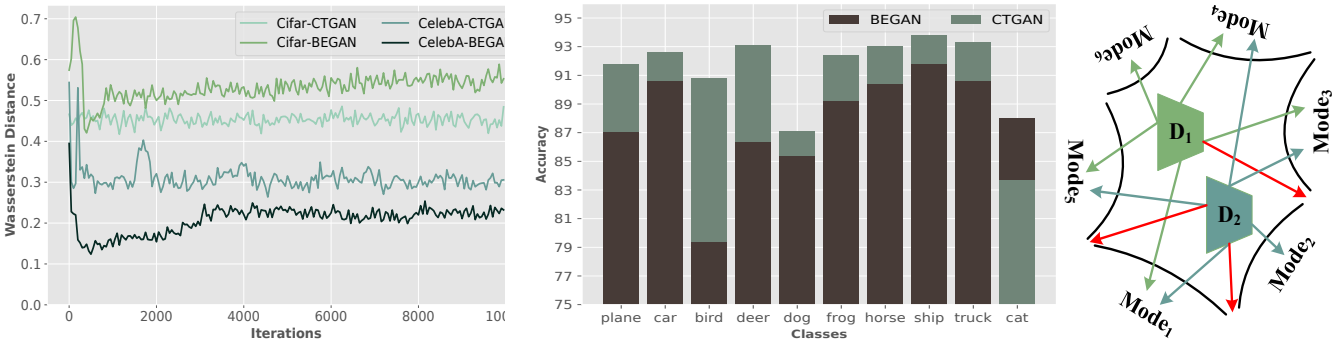
1 Introduction

Generative methods are one of the most promising approaches toward automatically learning features from a given high-dimension data distribution and then producing new samples approximate the truth [Goodfellow *et al.*, 2016]. It has been rapidly growing in recent years and successfully applied in a wide of tasks and applications. Currently, the most prominent approaches are followed as Variational Auto Encoders(VAEs) [Kingma *et al.*, 2014], Generative Adversarial Networks(GANs) [Goodfellow *et al.*, 2014] and a unifying framework combining the best of both like WAE [Tolstikhin *et al.*, 2018]. GANs among them have the most significant impact. Typically, it cast the generative modeling as a two-player game between two networks. A generator maps a random latent variable z to the data distribution $\mathbb{P}_r$ without calculating the sample likelihood, and a discriminator evaluates the quality of the generated samples by determining if a sample x belongs to $\mathbb{P}_r$ [Wei *et al.*, 2018; Berthelot *et al.*, 2017]. Finally, GANs can produce very visually appealing samples but easily suffer from the training instability and the mode collapse problems [Arjovsky *et al.*, 2017; Kodali *et al.*, 2017; Miyato *et al.*, 2018].

Many researchers have observed and tried to explain these drawbacks. First, gradients will point to more or less random directions if the data distribution and the generated distribution do not have substantial overlap and are too easy to tell apart. Furthermore, there would be a perfect discriminator that can distinguish the generated distribution from the target. The training will be completely stopped and resulting in unstable training [Arjovsky and Bottou, 2017; Karras *et al.*, 2018]. On the other hand, GANs always search for a discontinuous transportation map in the latent space of continuous mapping, which leads to mode collapse [Lei *et al.*, 2019; An *et al.*, 2020]. Although some works have been proposed to tackle these problems by introducing multiple discriminators or multiple generators, they don't seem to be clearing up.

Experiments with two popular cost functions in GANs are conducted to explore the reason for that, including CTGAN [Wei *et al.*, 2018] and BEGAN [Berthelot *et al.*, 2017]. As shown in Figure 1(a), the Wasserstein distance between the real distribution and the generated distribution of the CTGAN is always smaller than the one with BEGAN in the whole training procedure using the CIFAR-10 training dataset [Krizhevsky and Hinton, 2009]. Instead, the distance is always larger than the one with BEGAN on the CelebA dataset [Krizhevsky and Hinton, 2009], although the experiment setting of these two datasets is the same. This means the performance of CTGAN is better than the one of BEGAN on the CIFAR-10 dataset but is weaker on the CelebA dataset. Besides, we can do semi-supervised learning with a standard classifier by simply adding samples from the generated distribution to the dataset and labeling them with a new class $y = K + 1$ [Salimans *et al.*, 2016]. The unsupervised learning costs used in this way are CTGAN and BEGAN, respec-

*Contact Author



(a) Distance Curve (b) Accuracy for Different Classes (c) Distribuitions

Figure 1: The figure at the left is the Wasserstein distance curve between the real and generated distribution in the training procedure. Two cost functions(CTGAN and BEGAN) and two datasets(CIFAR-10 and CelebA datasets) are introduced in this experiment. The middle one is the accuracy of the semi-supervised learning method with these two cost functions on the CIFAR-10 dataset. Meanwhile, the last one at the right is the distribution in the data space of two discriminators with different cost functions. Arrows with different colors show the gradient of these two discriminators.

tively. As shown in Figure 1(b), although the total accuracy of the CTGAN one is significantly higher than the one with BEGAN, the accuracy for the class termed cat is poorer than the BEGAN one. These indicate the preference of diverse cost functions is different, and a single cost function cannot learn all modes. Map the distributions into a latent space and be shown in Figure 1(c). Any discriminator cannot search all modes in the latent space on its own, whether D_1 or D_2. In contrast, it is possible if simultaneously introducing all discriminators. Besides, there might be some errors(red arrows) in the latent space. Their gradients will move to the spurious samples rather than the data distribution, thus leading to the mode mixture problem, which means the generated samples look like a mixture of different modes [Lei *et al.*, 2019].

Inspired by these, we propose CES-GAN, a framework to tackle the mode collapse problem, which simultaneously trains multiple discriminators with different cost functions. More gradients point to diverse modes are yielded in this way, and thus we can search for more transportation maps in the latent space. Significantly, this means we can search all modes potentially. However, it may also lead to the generator being hard to train. According to the Cannikin Law, the shortest slab determines the capacity of a wooden tub. The gradient yielded from a weak discriminator will likely mislead the generator and degrade performance. Therefore, we have also proposed a gradient selecting mechanism to pick up proper gradients and speed the training up. Discriminators in CES-GAN are divided into two types: a core discriminator to produce template gradients and a group of auxiliary discriminators to provide supplementary gradients. Gradients yielded from the core discriminator are always used to update the generator. Instead, we calculate the resultant Wasserstein distance between the data distribution and the generated distribution and compare it with the latest smallest distance whenever we update the generator by descending the gradient yielded from an auxiliary discriminator. Only the gradient related to a smaller distance is the proper one to update the generator. Besides, not only useless gradients but an auxiliary discriminator will also be discarded if it cannot produce proper gradients. Finally, CES-GAN can avoid falling into the mode collapse problem benefits from gradient selection.

Mathematical statements are provided to prove our assumptions. Meanwhile, we conduct extensive experiments on various well-known datasets to evaluate the contributions. Results show that CES-GAN achieves excellent performance boosts and is less prone to mode collapse. This work provides these primary contributions:

- A novel framework named CES-GAN is proposed in this paper to tackle the mode collapse problem for GANs. To the best of our knowledge, it is the first work to introduce multiple discriminators with different cost functions, providing a new approach to solving the mode collapse problem.
- According to the Cannikin Law, the gradient yielded from a weak discriminator is likely to mislead the generator and even reduce the performance. Thus, we have also proposed a gradient selecting mechanism to avoid falling into this trap by selecting proper gradients.
- We provide mathematical statements to prove our assumptions. Besides, extensive experiments are conducted to verify the performance in terms of fighting against the mode collapse problem and contrast with other works.

2 Related Works

In recent years, many training methods with different tricks have been applied to deal with the mode collapse problem. Although the original GAN only involves training a single discriminator and an available generator, some attempts have begun to consider the framework of multiple components. First, some works have been made to train GANs with mixture generators to overcome the mode collapse problem. Tolstikhin et al. [Tolstikhin *et al.*, 2017] apply boosting techniques to train a mixture of generators by sequentially training and adding generators to the mixture. However, sequen-

tially training many generators is computationally expensive. Quan et al. [Hoang *et al.*, 2018] have proposed a framework that contains a set of generators, a discriminator, and a classifier. Generators create samples that are intended to come from the same distribution as the training data, whilst the discriminator determines whether samples are true data or generated by generators. The classifier specifies which generator a sample comes from. Then, there are some works to solve this problem using multiple discriminators. Nguyen et al. have proposed a novel framework with two discriminators to combine the Kullback-Leibler(KL) and reverse KL divergences into a unified objective function. Thus it exploits the complementary statistical properties from these divergences to effectively diversify the estimated density in capturing multimodes. Zhang et al. [Zhaoyu and Jun, 2019] have also proposed a similar architecture with two different discriminators named STDGAN, based on the ResBlock, Spectral Normalization, and Scaled Exponential Linear Units. Ishan Durugkar et al. [Durugkar *et al.*, 2017] have proposed a framework that extends GANs to multiple discriminators, named GMAN, allowing training with the original untampered minimax objective. As a result, the generator can automatically regulate the training and reach higher performance in a fraction of the training time required for the standard GAN model. Goncalo Mordido et al. [Mordido *et al.*, 2020] propose to tackle the mode collapse problem in GANs by using multiple discriminators and assigning a different portion of each minibatch, called micro-batch, to each discriminator. Finally, an alternative framework in which many GANs or their variants are trained simultaneously is proposed, but it is always expensive to calculate. Although many works have been offered to tackle the mode collapse problem, no work has introduced different cost functions and gradient selection.

3 Problem Statement

3.1 Motivation

GANs always want to search for a mapping from white noise to the latent space. Unfortunately, the searching will not converge or converge to one continuous branch of the target mapping, leading to the mode collapse problem. On the other hand, the modes in the latent space searched by various cost functions are different. They indicate we can potentially search all data modes in the latent space by simultaneously training multiple discriminators with different cost functions. Unfortunately, it may also lead to the generator being hard to train. Because the performance between discriminators is unbalanced, the weaker one among them could yield the wrong gradient can mislead the generator. According to the Cannikin Law, it even causes performance degeneration. Therefore, a gradient selecting mechanism is introduced to pick up proper gradients.

Similar to the mentioned experiment, we do semi-supervised learning with the proposed method. The total accuracy on the whole CIFAR-10 test set is 92.32%(+1.16%), and the accuracy for the class of cats is 87.9%(+4.2%). Remarkably, we can improve the classification performance in the cat class and the feature representation, thus improving the total accuracy. This is enabled by searching more data modes and preventing wrong transportation maps. We then show our mathematical statement and the sketch of proofs as follows.

3.2 Mathematical Statment

We first recall the regularity analysis for mode collapse [Lei *et al.*, 2019]. The generator maps $g_\theta : (\mathcal{Z}, \zeta) \longrightarrow (\sum, \mu_\theta)$ can be further decomposed into two steps, which can be described as follows.

$$g_\theta : (\mathcal{Z}, \zeta) \xrightarrow{T} (\mathcal{Z}, \mu) \xrightarrow{g} (\Sigma, \mu_\theta) \tag{1}$$

where T is a transportation map, maps the white noise ζ to μ in the latent space $\mathcal{Z}$, g is the manifold parameterization, maps local coordinates in the latent space to the manifold Σ. By manifold structure assumption, the local chart representation $g : \mathcal{Z} \longrightarrow \Sigma$ is continuous. However, according to the regularity theory of optimal transportation map, except in very rare situations, the transportation map T is always discontinuous. This intrinsic conflict will lead to the mode collapse problem.

Besides, following [Arjovsky and Bottou, 2017], the update of the generator will be completely stopped if there exists a perfect discriminator D^* that can perfectly distinguish the generated distribution from the target. Let data distribution $\mathbb{P}_r$ and generated distribution $\mathbb{P}_g$ be two distributions that have support contained in two closed manifolds M and P that don't perfectly align and don't have full dimension. Under this case, D^* is smooth and constant almost everywhere in $\mathcal{M}$ and $\mathcal{P}$ if both of their supports are disjoint or lie on low dimensional manifolds.

The discriminator is hard to search all transportation maps in the latent space, and some of the transportation maps searched by the discriminator are wrong. Meanwhile, the training dynamics are unstable that easily fall into the perfect discriminator trap. Instead, more transportation maps can be searched in the latent space by simultaneously training multiple discriminators with different cost functions, preventing wrong transportation maps and thus tackling the mode collapse problem. Besides, we can alleviate the perfect discriminator problem due to the mixed generated distribution and the data distribution is hard to be perfectly disjointed. We prove this in Appendix A.

4 Framework

In contrast to the conventional game involving one generator against a single adversary, multiple discriminators with different cost functions are trained to do the qualitative assessment in our framework. To this end, an analogy to a game among k discriminators and a generator can be formulated, defined as $D_{1:i:k}$ and G, respectively, as shown in Figure 2.

Discriminators are divided into two types: a core discriminator D_k to yield template gradients and a group of auxiliary discriminators $D_{1:i:k-1}$ to provide more gradients used as supplementary. Unlike the sequential training, the generator can be updated only once by descending its stochastic gradient from a discriminator in each training iteration. p is a given probability from 0 to 1 to control updating the generator. For example, we update the generator by descending the

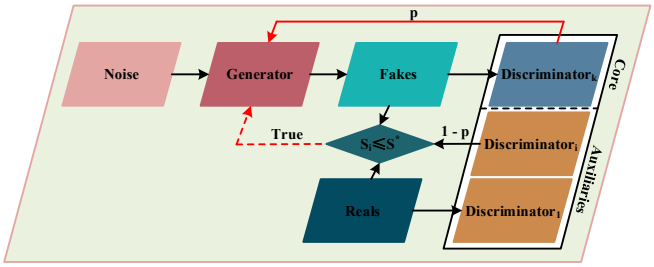


Figure 2: The proposed framework for CES-GAN, all components are parameterized by neural networks. Train the generator successfully by the core discriminator and auxiliary discriminators is shown with the solid red lines and the red dashed lines, respectively.

gradient from the core discriminator when the case of p happens. Otherwise, the generator can be updated by descending the gradient from one of the auxiliary discriminators. Thus, the cost function that the critic solves is:

$$\mathcal{L}_D = \frac{(1-p)}{k-1}\sum_{i=1}^{k-1}\mathcal{L}_{D_i} + p\mathcal{L}_{D_k} \tag{2}$$

However, it may also lead to the generator being hard to train because the performance between discriminators is unbalanced. According to the Cannikin Law, the shortest slab determines the capacity of a wooden tub, which indicates the gradient from a weak discriminator might be wrong to train the generator even mislead it. Thus, we have also proposed a gradient selecting mechanism in each training iteration to avoid falling into this trap. We show this procedure of a training iteration in the Algorithm 1.

The gradient yielded from the core discriminator is always used to update the generator. Instead, whenever to update the generator by descending the gradient from one of the auxiliary discriminators, we will calculate the resultant distance between the data distribution and the generated distribution. It will be used to compare with the latest smallest distance, only the gradient related to a smaller distance is regarded as the proper one to update the generator. In particular, the Wasserstein distance performs better than KL, JS, and TV divergence. It correlates with convergence and sample quality, which can serve as a useful metric over probability distribution [Gulrajani *et al.*, 2017; Salimans *et al.*, 2018]. Therefore, the Wasserstein distance is introduced to evaluate is the gradient proper to update the generator. Not only useless gradients, an auxiliary discriminator which has continuously failed to update the generator 200 times will also be discarded. This is because such a discriminator is hard to provide proper gradients and leads to a waste of computing resources. With the discarding of such auxiliary discriminators, the time of a training iteration can be decreased.

5 Experiments

We have conducted extensive experiments to evaluate the performance of CES-GAN, in which three well-known cost functions in GANs have been used: the standard objective function, BEGAN, and CTGAN. CTGAN, among them, is always used as the core discriminator. More options such as the number of discriminators, the choice of the core discriminator, and the combination of different cost functions will be discussed in Appendix B. Adam optimized with an initial learning rate of 1e-4 has been used to train them, and no transform or data augmentation has been utilized in these experiments. The prior noise input to the generator follows the random Gaussian distribution with 128 sizes. All the experiments are implemented and evaluated with Pytorch on 8× Nvidia Geforce GTX 1080 Ti [Paszke *et al.*, 2019].

The Inception Score(IS) [Salimans *et al.*, 2016] and the Fréchet Inception Distance(FID) [Dowson and Landau, 1982; Heusel *et al.*, 2017] are used for quantitative evaluation of image quality. Inception Score is a metric that computes the KL divergence between the conditional class distribution and the marginal class distribution, and Fréchet Inception Distance is a more principled and comprehensive metric that is more consistent with human evaluation in assessing the realism and variation of the generated samples [Zhang *et al.*, 2018]. Accuracy is a metric used for evaluating the performance of semi-supervised learning in this work, which is the ratio between the number of correct predictions and the total number of data points in the dataset.

The proposed framework is first evaluated on three small

Algorithm 1 The gradient selecting mechanism for CES-GAN in each training iteration

Input: The prior noise input to the generator follows the random Gaussian distribution with 128 sizes.
Parameter: The number of steps to apply to the discriminator n, the given probability p, and the latest smallest distance S^* between the data distribution and the generated distribution.

for n steps **do**
  Update all of discriminators
  if p **then**
    Update the generator using the gradient from D_k
    Calculate the distance between $\mathbb{P}_r$ and $\mathbb{P}_g$, as S_i
    if $S_i \leqslant S^*$ **then**
      $S_i \rightarrow S^*$
    end if
  else
    for $j = 1$ to $k-1$ **do**
      Update the generator using the gradient from D_j
      Calculate the distance between $\mathbb{P}_r$ and $\mathbb{P}_g$, as S_i
      if $S_i \leqslant S^*$ **then**
        $S_i \rightarrow S^*$
        Break
      else
        Reload the parameter for the generator
        if $j = k-1$ **then**
          return the case of p
        end if
      end if
    end for
  end if
end for

but well-known estimating the performance of combat the mode collapse problem datasets: the prevalent MNIST [LeCun *et al.*, 1998], the Stacked MNIST [Lin *et al.*, 2018], and the synthetic dataset. We compare with the SOTA methods both quantitatively and qualitatively on the CIFAR-10 dataset. Besides, Results on a large-scale dataset named CelebA [Liu *et al.*, 2015; Karras *et al.*, 2018] are described and compared with other works. Another large-scale dataset called ImageNet will be shown in Appendix D.

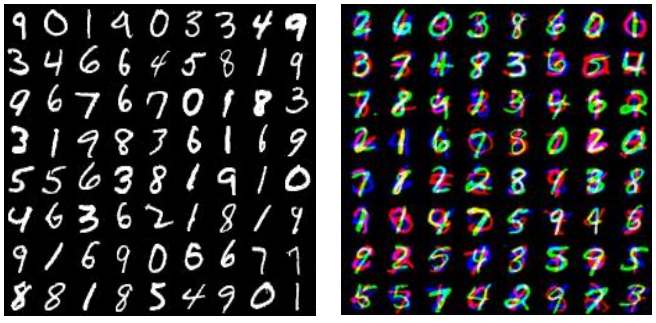
5.1 MNIST and Stacked MNIST Datasets

The MNIST of handwritten digits is a subset of a more extensive set available from NIST, which provides 70,000 examples in total, and 10,000 of them are left out for the testing. The digits have been size-normalized and centered in a fixed-sized image. Besides, the Stacked MNIST dataset is a dataset in which each image consists of three randomly selected MNIST images stacked into a three-channel image in RGB that has $10\times10\times10$=1,000 modes.

We use only 1,000 MNIST images to train the networks without any data augmentation. As Figure 3(a) shows, the contrasts between the foreground and the background are sharp to see. Following [Lin *et al.*, 2018], we then have also evaluated our performance on the Stacked MNIST dataset. The number of samples used to train the networks is 128,000 samples, with a batch size of 64 samples. We generate samples from the generator, and each of the three channels in each sample is classified by a pre-trained classifier to determine which of 1,000 modes the sample belongs to, as shown in Figure 3(b). The number of observed modes, as well as the KL divergence of the generated mode distributions, are measured for evaluation. Finally, CES-GAN is sufficient to fully capture all the modes in the benchmark test, and their empirical KL divergence is 0.04 ± 0.006 evaluated by 26,000 samples.

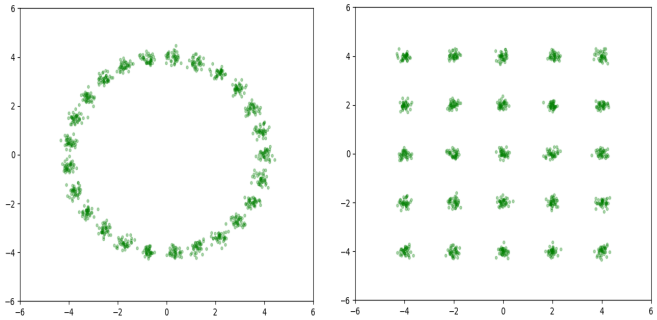
5.2 Toy Datasets

Since synthetic dataset consists of explicit distributions and known modes, mode collapse and the quality of the generated sample can be accurately measured. We use 2D-ring and 2D-grid with 25 modes to evaluate the number of modes, the percentage of high-quality samples, and the reverse Kullback-Leibler(KL) divergence, as shown in Figure 4. The number of modes counts the number of modes captured by samples produced in a generative model. The percentage of high-quality samples measures the proportion of samples generated within three standard deviations of the nearest mode. The Reverse KL divergence measures how well generated samples balance among all modes regarding the real distribution.



(a) MNIST (b) Stacked MNIST

Figure 3: MNIST and Stacked MNIST datasets



(a) 2D-ring (b) 2D-grid

Figure 4: Toy datasets

Under the standard benchmark settings in [Lin *et al.*, 2018], we train GANs with 100,000 total samples and a batch size of 100 pieces and report the average of 10 independent experiment results. For the 2D-ring dataset, CES-GAN achieves 24.8$\pm$0.2 of the modes number, 99.9%$\pm$0.1% of the percentage of high-quality samples, and 0.01$\pm$0.01 of the reverse KL divergence. Then, the mode number we have achieved on the 2D-grid dataset is 24.9$\pm$0.1, with a percentage of high-quality samples of 99.8%$\pm$0.2%, and a reverse KL divergence of 0.01$\pm$0.01. CES-GAN matches or outperforms SOTA schemes in terms of these three metrics, as shown in Appendix C. Therefore, we can say that the proposed method is effective for fighting against the mode collapse problem.

5.3 CIFAR-10 Dataset

CIFAR-10 is a publicly accessible dataset that contains 50,000 natural images that have been the most widely used for the image classification study and to test the performance of generative models [Krizhevsky and Hinton, 2009]. All images in this dataset are in color with an image size of 32×32 pixels. We first use only 1,000 images to train a small convolutional neural network and show the results in Figure 5(a). We achieve an inception score of 5.62 ± 0.16 and an FID of 35.87 ± 0.71, which is the state-of-the-art result to the best

(a) Generated by a small CNN (b) Generated by a ResNet

Figure 5: CIFAR-10 generated images without supervised by a small CNN and a large-scale ResNet, respectively.

Generative Methods	Inc-Score	FID
GMAN[Hoang *et al.*, 2018]	6.00	60.3
D2GAN[Nguyen *et al.*, 2017]	7.15	32.1
MADGAN[Ghosh *et al.*, 2018]	7.46	30.7
MGAN[Hoang *et al.*, 2018]	8.33	20.5
TDGAN[Zhaoyu and Jun, 2019]	8.34	21.4
CES-GAN	**8.48**	**19.7**

Table 1: The Inception Score(Inc-Score) and the Fréchet Inception Distance(FID) for different generative methods trained on the CIFAR-10 dataset.

of our knowledge. Then, we have trained another large-scale ResNet [He *et al.*, 2016] on the whole training set for an unsupervised task, as shown in Figure 5(b).

We can capture all modes in the data space in both cases, which proves the proposed method is effective for combating the mode collapse problem. Besides, CES-GAN achieves the highest inception score and the lowest Fréchet Inception Distance compared to other multi-components methods, as shown in Table 1.

For the semi-supervised learning approach, same with [Wei *et al.*, 2018], we follow the standard training/test split of the dataset but use only 4,000 labels in training. Regular data augmentation with flipping the images horizontally and randomly translating the images within -2 and 2 pixels is utilized. We report the semi-supervised learning results in Table 2. Compared to several very competitive methods, CES-GAN can achieve a state-of-the-art result that outperforms all the GAN-based methods.

5.4 CelebFaces Attributes Dataset

CelebFaces Attributes Dataset (CelebA) is a large-scale face attributes dataset with more than 200K celebrity images, each with 40 attribute annotations. The images in this dataset cover large pose variations and background clutter which indicates

Semi-Supervised Learning Methods	Accuracy
Ladder [Rasmus *et al.*, 2015]	20.40 ± 0.47
VAT [Miyato *et al.*, 2017]	10.55 ± 0.21
TE [Laine and Aila, 2016]	12.16 ± 0.24
Te-Student [Tarvainen and Valpola, 2017]	12.31 ± 0.28
GANs [Salimans *et al.*, 2016]	18.63 ± 2.32
Invariances [Abhishek *et al.*, 2017]	16.78 ± 1.80
CTGAN [Wei *et al.*, 2018]	9.98 ± 0.21
CES-GAN	$\mathbf{9.36 \pm 0.20}$

Table 2: The test error of semi-supervised learning approaches on the CIFAR-10 dataset.

Figure 6: Samples of fake images generated by the proposed method using the CelebA and the CelebA-HQ datasets, respectively.

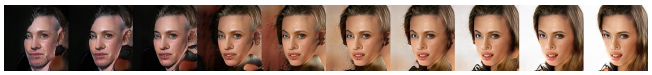


Figure 7: The linear interpolation from a man to a woman.

CelebA has large diversities, large quantities, and rich annotations. We trained networks for the resolution of 128×128, as shown in the left of Figure 6. Furthermore, we also test the proposed method on images with high resolution, namely the CelebA-HQ dataset with the image size to be 256×256, as shown in Figure 6 right. The FID of the proposed method on the CelebA dataset is 7.4 and 7.1 on the CelebA-HQ dataset. It is more photo-realistic than AE-OT-GAN(7.6 and 7.4, respectively) and has less computational cost.

These images in Figure 7 were not part of the training data. The first and last columns contain the images to be represented and interpolated. The images immediately next to them are their corresponding approximations, while the images in-between are the results of linear interpolation. The generated images look close to the real ones, and the interpolations show good continuity.

6 Conclusion

We have proposed a novel framework to tackle the mode collapse problem by simultaneously training multiple discriminators with different cost functions called CES-GAN. Unfortunately, it may also lead to the generator being hard to train because the performance between discriminators is unbalanced, according to the Cannikin Law. Therefore, a gradient selecting mechanism is also proposed to pick up proper gradients. We provide mathematical statements to prove assumptions and conduct extensive experiments to verify the performance. CES-GAN is able to produce more photo-realistic samples and is less prone to mode collapse.

References

[Abhishek *et al.*, 2017] Abhishek, Prasanna, and Fletcher. Improved semi-supervised learning with gans using manifold invariances. CoRR, abs/1705.08850, 2017.

[An *et al.*, 2020] An, Guo, Lei, Luo, Yau, and Gu. Ae-ot: A new generative model based on extended semi-discrete optimal transport. In *ICLR*, 2020.

[Arjovsky and Bottou, 2017] Arjovsky and Bottou. Towards principled methods for training generative adversarial networks. In *ICLR*, 2017.

[Arjovsky *et al.*, 2017] Arjovsky, Chintala, and Bottou. Wasserstein generative adversarial networks. In *ICML*, pages 214–223, 2017.

[Berthelot *et al.*, 2017] Berthelot, Thomas, and Luke. Began: Boundary equilibrium generative adversarial networks. CoRR, abs/1703.10717, 2017.

[Dowson and Landau, 1982] Dowson and Landau. The frechet distance between multvariate normal distributions. *Journal of Multivariate Analysis*, 12(3):450–455, 1982.

[Durugkar *et al.*, 2017] Durugkar, Gemp, and Mahadevan. Generative multi-adversarial networks. In *ICLR*, 2017.

[Ghosh *et al.*, 2018] Ghosh, Kulharia, Namboodiri, Torr, and Dokania. Multi-agent diverse generative adversarial networks. In *CVPR*, pages 8513–8521, 2018.

[Goodfellow *et al.*, 2014] Goodfellow, Pouget, Mirza, Xu, Warde, Ozair, Courville, and Bengio. Generative adversarial nets. In *NIPS*, 2014.

[Goodfellow *et al.*, 2016] Goodfellow, Bengio, and Courville. *Deep Learning*. MIT Press, 2016.

[Gulrajani *et al.*, 2017] Gulrajani, Ahmed, Arjovsky, Dumoulin, and Courville. Improved training of wasserstein gans. In *NIPS*, 2017.

[He *et al.*, 2016] He, Zhang, Ren, and Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[Heusel *et al.*, 2017] Heusel, Ramsauer, Unterthiner, Nessler, and Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *NIPS*, 2017.

[Hoang *et al.*, 2018] Hoang, Nguyen, Le, and Phung. Multi-generator generative adversarial nets. In *ICLR*, 2018.

[Karras *et al.*, 2018] Karras, Aila, Laine, and Lehtinen. Progressive growing of gans for improved quality, stability, and variation. In *ICLR*, 2018.

[Kingma *et al.*, 2014] Kingma, Diederik, and Welling. Auto-encoding variational bayes. In *ICLR*, 2014.

[Kodali *et al.*, 2017] Kodali, Abernethy, Hays, and Kira. How to train your dragan. CoRR, abs/1705.07215, 2017.

[Krizhevsky and Hinton, 2009] Krizhevsky and Hinton. Learning multiple layers of features from tiny images. *Handbook of Systemic Autoimmune Diseases*, 1(4), 2009.

[Laine and Aila, 2016] Laine and Aila. Temporal ensembling for semi-supervised learning. CoRR, abs/1610.02242, 2016.

[LeCun *et al.*, 1998] LeCun, Bottou, Bengio, and Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[Lei *et al.*, 2019] Lei, Guo, An, Qi, Luo, Yau, and Gu. Mode collapse and regularity of optimal transportation maps. CoRR, abs/1902.02934, 2019.

[Lin *et al.*, 2018] Lin, Khetan, Fanti, and Sewoong. Pacgan: The power of two samples in generative adversarial networks. In *NIPS*, 2018.

[Liu *et al.*, 2015] Liu, Luo, Wang, and Tang. Deep learning face attributes in the wild. In *ICCV*, pages 3730–3738, 2015.

[Miyato *et al.*, 2017] Miyato, Maeda, Koyama, and Ishii. Virtual adversarial training: A regularization method for supervised and semi-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):1979–1993, 2017.

[Miyato *et al.*, 2018] Miyato, Kataoka, Koyama, and Yoshida. Spectral normalization for generative adversarial networks. In *ICLR*, 2018.

[Mordido *et al.*, 2020] Mordido, Yang, and Meinel. microbatchgan: Stimulating diversity with multi-adversarial discrimination. In *WACV*, pages 3061–3070, 2020.

[Nguyen *et al.*, 2017] Nguyen, Le, Vu, and Phung. Dual discriminator generative adversarial nets. In *NIPS*, 2017.

[Paszke *et al.*, 2019] Paszke, Gross, Massa, Lerer, Bradbury, Chanan, Killeen, Lin, Gimelshein, and Antiga. Pytorch: An imperative style, high-performance deep learning library. In *NIPS*, 2019.

[Rasmus *et al.*, 2015] Rasmus, Berglund, Honkala, Valpola, and Raiko. Semi-supervised learning with ladder networks. In *NIPS*, 2015.

[Salimans *et al.*, 2016] Salimans, Goodfellow, Zaremba, Cheung, Radford, and Chen. Improved techniques for training gans. In *NIPS*, 2016.

[Salimans *et al.*, 2018] Salimans, Zhang, Radford, and Metaxas. Improving gans using optimal transport. CoRR, abs/1805.05573, 2018.

[Tarvainen and Valpola, 2017] Tarvainen and Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *NIPS*, 2017.

[Tolstikhin *et al.*, 2017] Tolstikhin, Gelly, Bousquet, Simon, and Schölkopf. Adagan: Boosting generative models. In *NIPS*, 2017.

[Tolstikhin *et al.*, 2018] Tolstikhin, Bousquet, Gelly, and Schoelkopf. Wasserstein auto-encoders. In *ICLR*, 2018.

[Wei *et al.*, 2018] Wei, Gong, Liu, Lu, and Wang. Improving the improved training of wasserstein gans: A consistency term and its dual effect. In *ICLR*, 2018.

[Zhang *et al.*, 2018] Zhang, Goodfellow, Metaxas, and Odena. Self-attention generative adversarial networks. In *ICML*, pages 7354–7363, 2018.

[Zhaoyu and Jun, 2019] Zhaoyu and Jun. Stdgan: Resblock based generative adversarial nets using spectral normalization and two different discriminators. In *ACM MM*, pages 674–682, 2019.