AN IMPROVED COMPOSITE FUNCTIONAL GRADIENT LEARNING BY WASSERSTEIN REGULARIZATION FOR GENERATIVE ADVERSARIAL NETWORKS

Anonymous authors
Paper under double-blind review

ABSTRACT

Generative adversarial networks (GANs) are usually trained by a minimax game which is notoriously and empirically known to be unstable. Recently, a totally new methodology called Composite Functional Gradient Learning (CFG) provides an alternative theoretical foundation for training GANs more stablely by employing a strong discriminator with logistic regression and functional gradient learning for the generator. However, the discriminator using logistic regression from the CFG framework is gradually hard to discriminate between real and fake images while the training steps go on. To address this problem, our key idea and contribution are to introduce the Wasserstein distance regularization into the CFG framework for the discriminator. This gives us a novel improved CFG formulation with more competitive generate image quality. In particular, we provide an intuitive explanation using logistic regression with Wasserstein regularization. The method helps to enhance the model gradients in training GANs to archives better image quality. Empirically, we compare our improved CFG with the original version. We show that the standard CFG is easy to stick into mode collapse problem, while our improved CFG works much better thanks to the newly added Wasserstein distance regularization. We conduct extensive experiments for image generation on different benchmarks, and it shows the efficacy of our improved CFG method.

1 INTRODUCTION

The GANs learn to sample random variable z from a known distribution p_z to approximate the real data distribution p_*. The minimax formulation is utilized by (Goodfellow et al., 2014) to optimize the generator and discriminator alternately. Typically, the generator is trained to synthesize samples by the learned underlying data distribution, while the discriminator differentiates the generated data from real data samples. Despite various training strategies (Arjovsky et al., 2017; Gulrajani et al., 2017) have been studied, GANs is still notoriously difficult to train due to its instability and the issues of mode collapse. Recently, the Composite Functional Gradient Learning (CFG) (Johnson & Zhang, 2019) has been presented as a novel theory for GANs which does not rely on the minimax formulation. The CFG reformulate the generator by the functional gradient learning. In training, the CFG employs the functional compositions to learn the generator greedily

$$G_t(z) = G_{t-1}(z) + \eta_t g(G_{t-1}(z)), (t = 1..., T) \tag{1}$$

Given the $G_0(z) \in \mathbb{R}^r$, the output $x = G_t(z)$ will approximate the data from distribution p_* with the learning rate η_t and the discrete time step from time $t = 1$ to time $t = T$. Each g function is estimated from the data. Accordingly, the discriminator works as a regressor to discriminate the data from real and the generator:

$$D \approx \underset{D}{\operatorname{argmin}} \mathcal{L}_R = \underset{D}{\operatorname{argmin}} \left[\mathbb{E}_{x\sim p_*} \ln\left(1 + e^{-D(x)}\right) + \mathbb{E}_{x\sim p_z} \ln\left(1 + e^{D(x)}\right)\right] \tag{2}$$

The CFG theory gives a new stable GANs method, and it shows that with a strong discriminator, the generator is more stable learned with functional gradient learning. However, empirically we found that the performance of CFG GANS is potentially very sensitive to the hyper-parameters, which may demand extraordinary efforts to tune parameters carefully. On the other hand, as a useful measurement of the distance between different probability distributions, the Optimal Transport Map theory is widely utilized to alleviate the mode collapse problem in GANs training. This motivates us to repurpose the Wasserstein distance from Optimal Transport Map theory to improve the efficacy of CFG framework further. For unified the name of the CFG framework, we called it ICFG in the rest of paper.

Formally, this paper presents an improved Composite Functional Gradient by enforcing the Wasserstein distance (ICFGW). Specifically, we incorporate the Wasserstein regularization into the Eq (3), which thus has bounded first derivatives. This results in a novel improved ICFG formulation with a better training process and image quality. We further give the intuitive explanation of employing Wasserstein regularization to enhance the model gradients for GANs. According to the updated ICFG theory, a new improved algorithm for learning generative adversarial models has been developed. We conduct extensive experiments for image generation on different benchmarks. We compare our ICFGW against the ICFG (Johnson & Zhang, 2019). We demonstrate that the Wasserstein regularization improves the efficacy of ICFG; and the ICFG with Wasserstein regularization improved the generated image quality. We summarize the main contributions as follows:

- For the first time, We introduce the Wasserstein regularization to the CFG framework, which trains GANs with the regressor and is susceptible to many hyper-parameters. The improved CFG learning formulation thus has better training process and image quality.
- We give an intuitive understanding of the reason why the origin CFG architecture with logistic regression has a weak effect on differentiating generated images from real images. The Wasserstein regularization will help learn the gradient of the network stably in training process.
- Empirically the experiments show that the ICFG-Network works better with Wasserstein regularization than the original CFG in MNIST, EMNIST, FashionMNIST, CIFAR10, SVHN and LSUN datasets. The new formulation gets a very competitive convergence speed and synthesizing results over the original CFG.

2 RELATED WORKS

2.1 GENERATIVE ADVERSARIAL NETWORKS

GANs alternately optimized a discriminative and a generative model in a min-max loss function (Goodfellow et al., 2014). There are various significant variants for image generation, such as DCGAN (Radford et al., 2015), Progressive Growing GAN (Karras et al., 2017), and BigGAN (Brock et al., 2018),SAGAN (Zhang et al., 2019), StyleGAN (Karras et al., 2019). In general, GANs are very difficult to be stably trained. Training GANs may suffer from various issues, including gradients vanishing, mode collapse, and so on (Che et al., 2016; Roth et al., 2017; Nowozin et al., 2016). Numerous excellent works have been made in addressing these issues. For example, WGAN (Arjovsky et al., 2017) and its extensions (Gulrajani et al., 2017; Nowozin et al., 2016; Mao et al., 2017; Li et al., 2017; Metz et al., 2016) uses the Wasserstein distance to improve the training stability of GANs. The normalization methods also contribute greatly to this issue (Ioffe & Szegedy, 2015; Miyato et al., 2018; Kurach et al., 2019; Zhang et al., 2020).

2.2 WASSERSTEIN DISTANCE AND APPLICATIONS IN GANS

The Optimal Transport Map (OTM) problem has been proposed by Monge. By using the linear program model and dual optimal method, Kantorovich has given a relaxation solution for the Optimal Transport Map. Wasserstein distance is the minimum cost function for the Kantorovich linear program model. The idea of Wasserstein distance has been introduced in WGAN(Arjovsky et al., 2017)(Gulrajani et al., 2017). Unlike the GANs working with JS-divergence or least square loss function always meeting the vanishing gradient problem, the Wasserstein distance could avoid the vanishing gradient of the discriminator. It also controls the diversity of fake images from the generator. The mechanism and regularization of applying Wasserstein distance in GANs have been widely explored in the following work. It has become a standard loss function in the GANs training(Zhang

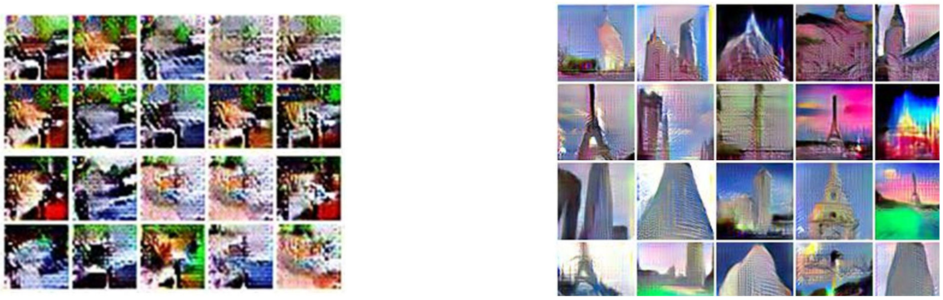


Figure 1: Synthesized images. The generated image from icfg is blur. The season is the weakness ability of the discriminator to differentiate low quality generated images the and the real images.

et al., 2019) (Karras et al., 2019)(Shen et al., 2021). On the other hand, few efforts are made on applying Wasserstein distance in regression or the classification. Some works provide theoretical proof that Wasserstein distance is helpful for the regression and the classification task (Jiang et al., 2020)(Frogner et al., 2015). (Frogner et al., 2015) develop a loss function for multi-label learning, based on the Wasserstein distance. They give a statistical learning bound for the loss; and describe an efficient learning algorithm based on this regularization. (Jiang et al., 2020) propose an approach to fair classification that enforces independence between the classifier outputs and sensitive information by minimizing Wasserstein-1 distances.

Inspired by these works, Wasserstein distance has been for the first time introduced into ICFG discriminator of framework; we theoretically show that the regression with Wasserstein regularization has a convergence with the methodology (Du et al., 2018)(Du et al., 2019). Meanwhile, we empirically show that such an idea can significantly improve the ICFG formulation.

3 IMPROVED ICFG WITH WASSERSTEIN REGULARIZATION

3.1 REGRESSION IN ICFG FRAMEWORK

We follow the definition in the original ICFG frameworkJohnson & Zhang (2019). This paper denotes the ICFG with a regression classifier as discriminator (Eq. 2) and a functional gradient learning generator (Eq. 1). The discriminator works as a regression classifier to discriminate the image from real and the generator, while the generator serves as the derivation function transforming the $G_t(Z)$ approximating the p_*, the density of real data. The sequence of transformation in Eq. 1 takes the discrete steps from time $t-1$ to t ($t = 1, \cdots, T$). Empirically, we found that the performance of GANs from ICFG framework is very sensitive to so many hyper-parameters, such as T, N, η_t and g as in Eq. 1. Particularly, the g is computed from data as

$$g(x) = s(x)\tilde{r}(x)f^{''}(\tilde{r}(x))\nabla D(x)$$

where $s(x)$ is an arbitrary scaling factor; $\tilde{r}(x) = e^{-D(x)}$. We define $\delta = s(x)\tilde{r}(x)f^{''}(\tilde{r}(x))$, and the image quality generated is very sensitive the value of δ. Furthermore, the training process of GANs by ICFG is much sensitive to the learning rate than that learned by minimax optimization. For example, with slightly changing the value of learning rate, the quality of synthesized images is changed from very high to very noisy, as empirically visualized in Fig. 1. In these figures, the lr is set as 0.0025 above and 0.000025 below respectively; and δ is 1 and 0.1 for SVHN and CIFAR10 individually. For example, if we set the lr to 0.0025, the results of SVHN will become very noisy. Furthermore, when we train the ICFG for more epochs with less tuned hyper-parameters, the model may still be inclined to get collapsed. For example, we show the blurred and noisy images in Fig. 1 . We refer these problem as the weakness ability of the discriminator to differentiate low quality generated images the and the real images. With different δ and lr, the gradient from the discriminator is too low to guide the generator update its weights. As a result, We should give a more stronger regularization in the discriminator.

3.2 Wasserstein regularization for the ICFGW

To resolve above problems, we introduce Wasserstein regularization into the ICFG framework. We will give an insight explanation of the Wasserstein regularization and empirical evaluations in the next two sections. Wasserstein regularization is a conception from the Wasserstein distance of Optimal Transport Map. The formulation of origin Wasserstein distance is Eq. (3), it stands for a minimum transport map from p to q. The $\gamma\left(x,y\right)$ stands for the Joint probability density distribution of p, q. $d\left(x,y\right)$ stands for the distance between p and q.

$$\mathcal{W}\left[p,q\right]=\inf_{\gamma\in\left[p,q\right]}\iint\gamma\left(x,y\right)d\left(x,y\right)\mathrm{d}x\mathrm{d}y \tag{3}$$

The formulation Eq. (3) is hard to compute, so the Kantorovich gives a duality form from the origin Optimal Transport Map using Linear Programming (LP). The problem also changes to Optimal Transport plan, the cost function is refered in Eq. (4). The $\int\left[p\left(x\right)f\left(x\right)-q\left(x\right)f\left(x\right)\right]\mathrm{d}x$ stands for the cost of the transport plan and $\left\|f\left(x\right)-f\left(y\right)\right\|\leq d\left(x,y\right)$ stands for the distance measurement. The duality form has a maximum optimization direction opposed from the origin minimum optimization direction.

$$\mathcal{W}\left[p,q\right]=\max_{f}\left\{\int\left[p\left(x\right)f\left(x\right)-q\left(x\right)f\left(x\right)\right]\mathrm{d}x\middle|\left\|f\left(x\right)-f\left(y\right)\right\|\leq d\left(x,y\right)\right\} \tag{4}$$

In the experiments, the Eq. (4) always be written as the discrete form like Eq. (5). The $\mathbb{E}_{x\sim p_*}$ stands for sampling from the real data and $\mathbb{E}_{x\sim p_z}$ stands for sampling from the generated data.f and D both stands for the discriminator and the G stands for the generator.

$$D_W=\underset{\left\|D\right\|_L\leq1}{\operatorname{argmax}}\ \mathcal{L}_W=\underset{\left\|D\right\|_L\leq1}{\operatorname{argmax}}\ \mathbb{E}_{x\sim p_*}\left[D\left(x\right)\right]-\mathbb{E}_{x\sim p_z}\left[D\left(G\left(z\right)\right)\right] \tag{5}$$

3.3 loss function for the Discriminator of ICFGW

As to implement the Wasserstein regularization to the discriminator of ICFG, we give a new loss function to replace the origin logistic regression loss. The formula of it is in the Eq. (6). It stands for a Convex combination of logistic regression and the Wasserstein distance. The parameter α is between 0 and 1. The symbol $\mathcal{L}_R$ stands for the loss function of ICFG discriminator from Eq. (2). $\mathcal{L}_W$ stands for the loss function of the Wasserstein distance from Eq. (5). The origin logistic regression is to find a hyper-plane to differentiate the different classes data clearly. But in the high-dimensional image space, it is difficult to find such a hyper-plane to differentiates the generated images from the real images clearly. As to resolve this problem, we add a Wasserstein regularization to the logistic regression loss function. With Wasserstein distance definition, the new hyper-plane will differentiates the generated images and real images in a maximum probabilistic distance. The new hyper-plane not only differentiates the data from generated or real images, but also gives a maximum probabilistic distance between generated and real images.

$$\mathcal{L}=\alpha\mathcal{L}_R+(1-\alpha)\mathcal{L}_W \tag{6}$$

3.4 Evaluations of update function for generator

This section will evaluate the $g(x)$ in the empirical process. $g(x)$ is a symbol from formula Eq. (3.1). It is a very import parameter for the ICFG because it controls the gradient of discriminator used to update the generated images. In the experiment, we use cfg-eta replace the theoretical $g(x)$. In the origin ICFG, the little change of cfg-eta such as change 0.25 to 1 will influence the training process and generated image quality. But in our ICFGW, the cfg-eta will be set to 10 or 100 as the change of our Wasserstein regularization parameter α. The reason for that is the Wasserstein regularization Eq .(5) and logistic regression Eq .(2) has a adversarial optimization direction. The logistic regression is to obtain a minimum discriminator but the Wasserstein regularization is to obtain a maximum distance between the generated images and real. In the ICFGW's train, the gradient from discriminator will be much smaller than the ICFG because of the adversarial progress. That is why cfg-eta value in ICFGW will be 10 or 100 times bigger than the origin ICFG.

NAME	DESCRIPTION	VALUE
B	training data batch size	64
U	discriminator update per epoch	1
N	examples for updating G per epoch	640
T	number of iterations in ICFG	25
LAMDA	a hyper-parameters for L constant	0.1
α	a hyper-parameters for Wasserstein regularization	0.9

Table 1: meta-parameters.

DATASET	LR	CFG-ETA	LAMDA
E/Fashion/MNIST	2.5e-4	1/10	0.1/None
SVHN	2.5e-4	1/10	0.1/None
CIFAR	2.5e-4	1/10	0.1/None
LSUN	2.5e-4	1/10/100	0.1/None

Table 2: learning rate and others parameters for experiment.None stands for no implementation of parameter. The cfg-eta value is very different from the ICFG. We gives a insight explanation in the methodology. LMADA value of None stands for don not use this parameter

4 Experiment

4.1 Datasets

We used MNIST(LeCun et al., 1998), ,FashionMNIST(Xiao et al., 2017), EMNIST(Cohen et al., 2017), CIFAR10(Krizhevsky et al., 2009), the Street View House Numbers dataset (SVHN)(Netzer et al., 2011), and the large-scale scene understanding (LSUN) dataset(Yu et al., 2015). We almost follow the origin ICFG choosing for the datasets. We also give some addition datasets for our experiments. These datasets are provided with class labels (digits '0' – '9' for MNIST, FashionMNIS, EMNIST and SVHN and 10 scene types for LSUN). A number of studies have used only one LSUN class ('bedroom') for image generation. The origin paper employs a balanced two-class dataset using the same number of training images from the 'bedroom' class and the 'living room' class (LSUN BR+LR)and a balanced dataset from 'tower' and 'bridge' (LSUN T+B). But we choose to use the 'bedroom'(LSUN B) and the LSUN Tower(LSUN T) to make our experiment because the origin ICFG performance not very well in these database and we also present the result of the (LSUN T+B) and (LSUN BR+LR) in our supplements.

4.2 Implementation Details

As to fairy compare to the origin ICFG, we take the same implement settings. All the experiments were done using a single NVIDIA Tesla v100 or a single NVIDIA RTX 2080TI. The meta-parameter values for ICFG were fixed to those in Tab. 1 unless otherwise specified. ICFG is sensitive with the setting of step size η for the generative image updated in order to approximate an appropriate value. In our work, we give the same lr setting with the ICFG but different η size. For example, we could utilize the η value $10x$ or $100x$ more or less than those in our ICFGW as to approximate an appropriate image quality. This can also achieve very good result. Wasserstein regularization α is a hyper-parameter for the experiment; we set the meta-parameter LAMDA 0.9. The base setting is presented in the Tab. 1 and Tab. 2. The cfg-eta and lr in Tab. 2 stands for the $\delta = s(x)\tilde{r}(x)f^{''}(\tilde{r}(x))$ and η symbol. In order to keep the theoretical symbol and our experiment code consistent, we use cfg-eta , lr and cfg-alpha stands for the theoretical symbol in the experiment section. In our experiment, we will make a little change with the lr and other hyper-parameters to compare the effect between ICFG and our methods.

4.3 Baselines

As a representative of comparison methods, we tested WGAN with the gradient penalty (WGANgp), the Least square GAN. Both of them always been the baseline in other GANs Network. The network

	WGAN	**LSGAN**	**ICFG**	**ICFGW**
MNIST	0.781	0.679	1.15	2.32
SVHN	0.913	0.87	1.39	2.65
CIFAR	3.53	3.41	4.02	4.45
LSUN B	2.382	2.312	3.046	3.029
LSUN T	3.67	3.52	4.428	4.92

Table 3: Inception Score Result.

architecture is the same as the ICFG to fairy comparison. The network architecture is composite of two types, the first is DCGAN and the other is resnet. The DCGAN is used for MNIST, FashionMNIST, EMNIST, CIFAR10,and SVHN. The resnet is used for the LSUN datasets.

4.4 Evaluation metrics

Generative adversarial models is known to be challenge to make reliable likelihood estimates. So we instead evaluated the visual quality of generated images by adopting the inception score Salimans et al. (2016) and Fréchet inception distanceHeusel et al. (2017). The intuition behind inception score is that high-quality generated images should lead to close to the real image. And the Fréchet inception distance indicate that the similarity between the generated images and the real image. We note that the inception score is limited, e.g., it would not detect mode collapse or missing modes. Apart from that, we found that it generally corresponds well to human perception.

In addition, we used Fréchet inception distance (FID) of. FID measures the distance between the distribution of $f(x_*)$ for real data $x_*$and the distribution of $f(x)$ for generated data x, where function f is set to convert an image to the internal representation of a classifier net- work; One advantage of this metric is that it would be high (poor) if mode collapse occurs, and a disadvantage is that its computation is relatively expensive.

In the results below, we call these two metrics the (inception) score and the Fréchet distance.

4.5 Result for our experiment

In this section, we present our Wasserstein regularization for ICFG experimental comparisons with others GAN-model approaches. We also present the result of our model with the approximate learning-rate, cfg-eta, cfg-alpha compare to the origin-paper and evaluate that our method archives better result in the different meta-parameters.

4.5.1 Inception Score results

We can see the result of Inception Score value among different GANs in the Tab. 3 and Tab. 4. Note that the IS scores are affected by many factors, we recompute all the IS scores in all our experiments with our local compute environments. As we measure, the Inception Score for the real data in the datasets are 2.58(MNIST), 9.56(CIFAR10), 4.62(SVHN), 4.78(LSUN T), 3.72(LSUN B), 3.72(LSUN B+L), 3.79(LSUN T+B), 5.8(LSUN C), 4.34(FashionMNIST), 2.2(EMNIST). We get the Inception Score function from this url[1]. There are very different from the ICFG experiment. So we choose to use our experiment IS value to compare the quality between different model. We can still find the same conclusion in different IS score via the relation between them. Tab. 3 and Tab. 4 presents the result of the IS score of every model. We can clearly find that the ICFG score and ICFGW score is very close, and the WGAN and LSGAN performance not so much good in all the datasets. Although the score is different from the ICFG, the relative relationship also stands for that ICFGW archives a better image quality than ICFG in the same database.

4.5.2 Fréchet Distance results.

We can see the result of the image quality measured by the Fréchet Distance score in relation to training time and image quantity. We compute the the Fréchet Distance with 20k generative images

[1] https://github.com/sbarratt/inception-score-pytorch

	ICFG	**ICFGW**
EMNIST	2.1	2.14
FashionMNIST	4.21	4.16
LSUN C	3.17	3.09
LSUN B+L	3.44	3.49
LSUN T+B	5.11	4.92

Table 4: Inception Score Result.

	WGAN	**LSGAN**	**ICFG**	**ICFGW**
MNIST	4.72	4.93	4.49	3.688
SVHN	5.87	5.87	5.53	5.9
CIFAR	36.24	36.24	27.89	27.51
LSUN B	18.72	18.72	12.25	11.8
LSUN T	22.76	22.76	16.43	20.86

Table 5: Fréchet Distance results.

and 20k real images from datasets. The same as IS score, We keep the same strategy of recomputing all the FID score in our local compute environments for a fair comparison. The codes of FID functions are in[2]. We compute the value of FID in our environments but the scores gets much higher than the value of ICFG paper. Although we use the same ICFG method to generate the image, the result is no so much different. So we will compare the FID score which is calculate in local environment with different model and datasets. We can see the FID result in the Tab. 5 and Tab. 6. ICFGW archives the best in the MNIST, Fashion MNIST, EMNIST, SVHN, CIFAR10 and the LSUN T, T+B, B+L datasets. The ICFG works the best in the LSUN B. The WGAN and the LSGAN still have not well performance. The reason is that, for a fair comparison with other methods, we do not use tuning tricks, and these methods are also sensitive to varying hyper-parameters.

The reason for that would like be we keep the origin network settings of both network and do not add much training tricks to it. Besides we do not tune the hyper-parameters very carefully. For we keep the same network as the ICFG and still get the best scores in the nine out of ten datasets.The FID result shows that our ICFGW is effective and helpful to generate high quality images.

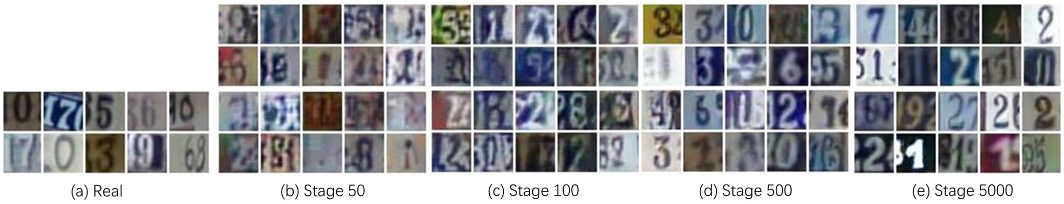
4.5.3 VISUAL INSPECTION OF GENERATED IMAGES

At last,we can see the image generate by different lr and cfg-eta and cfg-alpha parameter. Figure 2,3,4,5,6,7,8 shows the result of the experiment with ICFGW and ICFG. We set the same lr in ICFGW and ICFG separately. Besides we present the same epoch generate image together to show our ICFGW method can archive the same or better effect as the ICFG method. In some datasets, the ICFG image has already collapsed while ICFGW still works very well.

[2] https://github.com/mseitzer/pytorch-fid

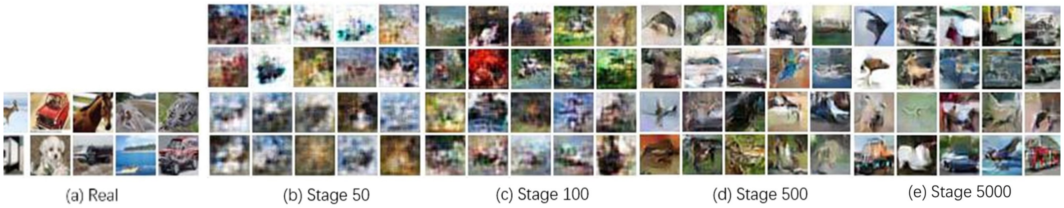
	ICFG	**ICFGW**
EMNIST	2.312	2.02
FashionMNIST	6.16	6.026
LSUN C	11.04	12.0
LSUN B+L	11.9	17.71
LSUN T+B	16.74	27.49

Table 6: Fréchet Distance results.



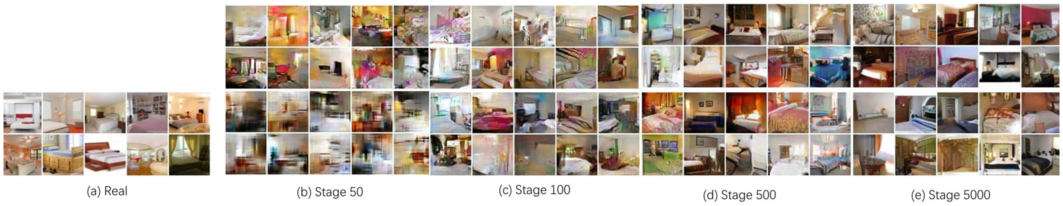
(a) Real (b) Stage 50 (c) Stage 100 (d) Stage 500 (e) Stage 5000

Figure 2: Result for SVHN:above is ICFGW and below is ICFG lr=0.00025 cfg-eta=1 cfg-alpha=0.9



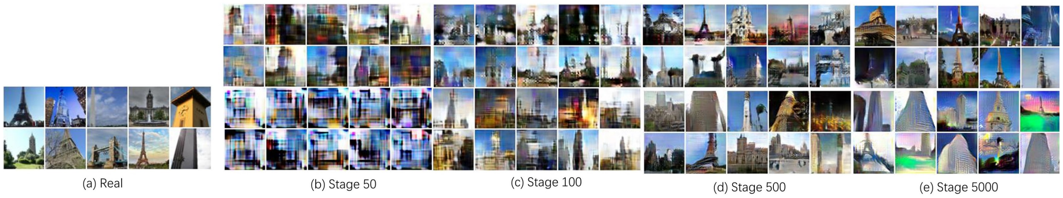
(a) Real (b) Stage 50 (c) Stage 100 (d) Stage 500 (e) Stage 5000

Figure 3: Result for CIFAR10:above is ICFGW and below is ICFG lr=0.00025 cfg-eta=0.5 cfg-alpha=0.9



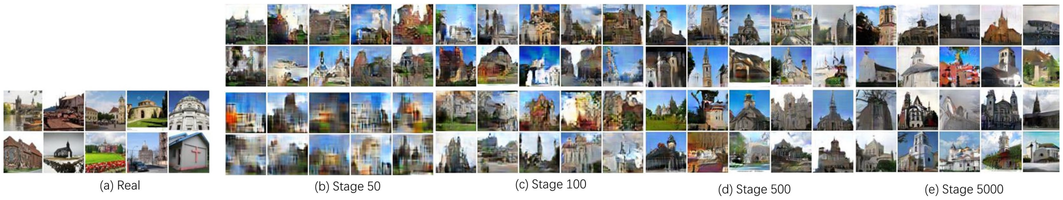
(a) Real (b) Stage 50 (c) Stage 100 (d) Stage 500 (e) Stage 5000

Figure 4: Result for LSUN B:above is ICFGW and below is ICFG lr=0.00025 cfg-eta=10 cfg-alpha=0.9



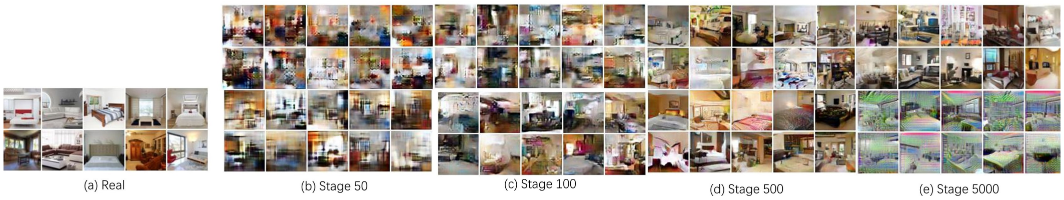
(a) Real (b) Stage 50 (c) Stage 100 (d) Stage 500 (e) Stage 5000

Figure 5: Result for LSUN T:above is ICFGW and below is ICFG lr=0.00025 cfg-eta=0.5 cfg-alpha=0.9



(a) Real (b) Stage 50 (c) Stage 100 (d) Stage 500 (e) Stage 5000

Figure 6: Result for LSUN C:above is ICFGW and below is ICFG lr=0.00025 cfg-eta=10 cfg-alpha=0.9



(a) Real (b) Stage 50 (c) Stage 100 (d) Stage 500 (e) Stage 5000

Figure 7: Result for LSUN B+L:above is ICFGW and below is ICFG lr=0.00025 cfg-eta=10 cfg-alpha=0.9

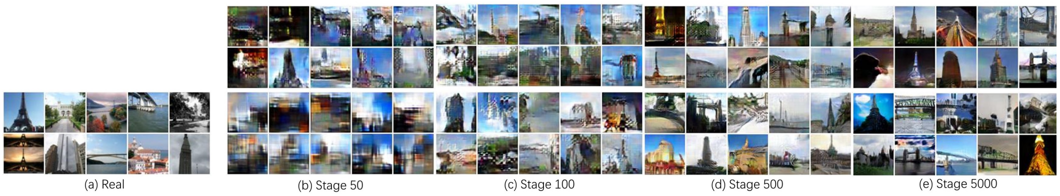


Figure 8: Result for LSUN T+B:above is ICFGW and below is ICFG lr=0.00025 cfg-eta=10 cfg-alpha=0.9

5 Conclusion

In this paper, we introduced the Wasserstein regularization into the Composite Functional Gradient Learning (CFG) which is a new theoretical way to train GAN. While the discriminator of standard ICFG is very sensitive to varying hyper-parameters. The effect of its differentiates is not good enough. But our ICFGW work much better with various hyper-parameters. The experiments results demonstrate that the proposed approach shows more stable performance compared with ICFG and other methods. The Inception score, Fréchet Distance, and the visual quality of generated image show that our method is more stable. In future work, we plan to investigate the generator object function from KL diversity to the Wasserstein distance as to achieve more stable and efficient GANs architecture.

References

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

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.

Gregory Cohen, Saeed Afshar, Jonathan Tapson, and Andre Van Schaik. Emnist: Extending mnist to handwritten letters. In *2017 International Joint Conference on Neural Networks (IJCNN)*, pp. 2921–2926. IEEE, 2017.

Simon Du, Jason Lee, Haochuan Li, Liwei Wang, and Xiyu Zhai. Gradient descent finds global minima of deep neural networks. In *International Conference on Machine Learning*, pp. 1675–1685. PMLR, 2019.

Simon S Du, Xiyu Zhai, Barnabas Poczos, and Aarti Singh. Gradient descent provably optimizes over-parameterized neural networks. *arXiv preprint arXiv:1810.02054*, 2018.

Charlie Frogner, Chiyuan Zhang, Hossein Mobahi, Mauricio Araya-Polo, and Tomaso Poggio. Learning with a wasserstein loss. *arXiv preprint arXiv:1506.05439*, 2015.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron Courville. Improved training of wasserstein gans. *arXiv preprint arXiv:1704.00028*, 2017.

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. *Advances in neural information processing systems*, 30, 2017.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pp. 448–456. PMLR, 2015.

Ray Jiang, Aldo Pacchiano, Tom Stepleton, Heinrich Jiang, and Silvia Chiappa. Wasserstein fair classification. In *Uncertainty in Artificial Intelligence*, pp. 862–872. PMLR, 2020.

Rie Johnson and Tong Zhang. A framework of composite functional gradient methods for generative adversarial models. *IEEE transactions on pattern analysis and machine intelligence*, 43(1):17–32, 2019.

Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4401–4410, 2019.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

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. PMLR, 2019.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Chun-Liang Li, Wei-Cheng Chang, Yu Cheng, Yiming Yang, and Barnabás Póczos. Mmd gan: Towards deeper understanding of moment matching network. *arXiv preprint arXiv:1705.08584*, 2017.

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.

Luke Metz, Ben Poole, David Pfau, and Jascha Sohl-Dickstein. Unrolled generative adversarial networks. *arXiv preprint arXiv:1611.02163*, 2016.

Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. *arXiv preprint arXiv:1802.05957*, 2018.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. 2011.

Sebastian Nowozin, Botond Cseke, and Ryota Tomioka. f-gan: Training generative neural samplers using variational divergence minimization. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, pp. 271–279, 2016.

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

Kevin Roth, Aurelien Lucchi, Sebastian Nowozin, and Thomas Hofmann. Stabilizing training of generative adversarial networks through regularization. *arXiv preprint arXiv:1705.09367*, 2017.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *Advances in neural information processing systems*, 29: 2234–2242, 2016.

Chengchao Shen, Youtan Yin, Xinchao Wang, Xubin Li, Jie Song, and Mingli Song. Training generative adversarial networks in one stage. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3350–3360, 2021.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms, 2017.

Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015.

Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. Self-attention generative adversarial networks. In *International conference on machine learning*, pp. 7354–7363. PMLR, 2019.

Zhihong Zhang, Yangbin Zeng, Lu Bai, Yiqun Hu, Meihong Wu, Shuai Wang, and Edwin R Hancock. Spectral bounding: Strictly satisfying the 1-lipschitz property for generative adversarial networks. *Pattern Recognition*, 105:107179, 2020.