

Article

CDL-GAN: Contrastive Distance Learning Generative Adversarial Network for Image Generation

Yingbo Zhou ^{1,2} , Pengcheng Zhao ^{1,2}, Weiqin Tong ¹ and Yongxin Zhu ^{2,*}

¹ School of Computer Engineering and Science, Shanghai University, Shanghai 200444, China; zhouyingbo2018@sari.ac.cn (Y.Z.); zhaopengcheng2018@sari.ac.cn (P.Z.); wqtong@shu.edu.cn (W.T.)

² Shanghai Advanced Research Institute, Chinese Academy of Sciences, Shanghai 201210, China

* Correspondence: zhuyongxin@sari.ac.cn

Abstract: While Generative Adversarial Networks (GANs) have shown promising performance in image generation, they suffer from numerous issues such as mode collapse and training instability. To stabilize GAN training and improve image synthesis quality with diversity, we propose a simple yet effective approach as Contrastive Distance Learning GAN (CDL-GAN) in this paper. Specifically, we add Consistent Contrastive Distance (CoCD) and Characteristic Contrastive Distance (ChCD) into a principled framework to improve GAN performance. The CoCD explicitly maximizes the ratio of the distance between generated images and the increment between noise vectors to strengthen image feature learning for the generator. The ChCD measures the sampling distance of the encoded images in Euler space to boost feature representations for the discriminator. We model the framework by employing Siamese Network as a module into GANs without any modification on the backbone. Both qualitative and quantitative experiments conducted on three public datasets demonstrate the effectiveness of our method.

Keywords: generative adversarial networks; contrastive distance learning; consistent contrastive distance; characteristic contrastive distance



Citation: Zhou, Y.; Zhao, P.; Tong, W.; Zhu, Y. CDL-GAN: Contrastive Distance Learning Generative Adversarial Network for Image Generation. *Appl. Sci.* **2021**, *11*, 1380. <https://doi.org/10.3390/app11041380>

Academic Editor:

Ángel González-Prieto

Received: 21 December 2020

Accepted: 29 January 2021

Published: 3 February 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Generative Adversarial Networks (GANs) [1] have shown incredible success as effective data-driven models for image synthesis, but have become exposed to inevitable training obstacles. To find the theoretic Nash equilibrium with non-convex objective functions, GAN needs to exploit image information from a continuous and high-dimensional parameter space. Because GAN training is substantially more complicated than a standard neural network, it is challenging to keep the training stable. As a result, the outputs of a generative model frequently become uncontrollable and are of poor quality. To handle these challenges, many solutions have been introduced to improve GANs' performance.

In recent times, numerous proposals for better designs and optimization of basic GANs have been reported. Mirza and Osindero [2], Huang et al. [3], Odena et al. [4] proposed a re-engineered network architecture based on conditional generation. Conditional GANs (cGANs) learn a conditional probability distribution from auxiliary information about real data. Wang et al. [5], Hoang, Quan et al. [6], and Nguyen Tu et al. [7] modeled generative–discriminative network pairs to increase the generation capacity of the generator. With multiple generators or discriminators, GANs can get more constructive gradient signals to learn intermediate representation. Larsen et al. [8], Makhzani et al. [9], Dumoulin et al. [10], Wang, Xiaoqing [11], and Kwak, Jeong gi et al. [12] use the most common encoder–decoder architecture to learn image features from latent space. These hybrid models are useful for addressing mode collapse. In the face of the oscillation in model parameters, previous work by Martin Arjovsky et al. [13], Takeru Miyato et al. [14], Salimans et al. [15], and Chunliang Li et al. [16] utilized appropriate loss functions to tackle

stability issues. Some added interventions such as normalization [13,14] and regularization [17,18] to the discriminator [19], the generator [20] or both together [21]. Others found new probability distance metric [15,16,22] to replace JS divergence. Concerning the optimization algorithm, several researchers proposed the use of another gradient descent optimization technique [23,24], or modified the training technique [25,26]. Within the solutions mentioned above, diverse characteristic GANs variants have made a difference to some extent in the GANs literature, but some issues still remain unsolved. In practice, these methods more or less result in poor image quality due to training instability or mode collapse.

In this work, we undertake a comprehensive and effective approach using Contrastive Distance Learning (CDL) to make GANs perform better. Motivated by Improved Consistency Regularized GAN (ICR-GAN) [21] and Mode Seeking GAN (MS-GAN) [20], we present Consistent Contrastive Distance (CoCD) to modulate the sensitivity of the generator with prior changes in the noise. In light of the work of Ansari [22] and Miyato [27], we establish additional Characteristic Contrastive Distance (ChCD) to capture more informative image features for the discriminator. The CoCD is aimed at mitigating mode collapse to improve the training stability, while the ChCD forces the discriminator to remember more useful high-level semantics to further improve image synthesis quality. In particular, to alleviate computational cost with two additional auxiliary losses, we design our framework with the Siamese modules.

In our experiments, we conduct comparisons on CDL-GAN across the existing optimized GAN models for three public datasets. CDL-GAN yields state-of-the-art image synthesis results among the existing models. In extensive qualitative and quantitative studies, we show that our work offers multi-faceted improvements. It achieves lower Fréchet Inception Distance (FID) [28] scores under the same training and evaluation conditions with different datasets. Meanwhile, it works well across a large range of GAN models with different hyperparameter sets of the Adam optimizer. Furthermore, our proposed approach further mitigates mode collapse and training instability issues in both the generator and discriminator.

In brief, the main contributions of this work can be summarised as follows in three fold:

- We propose a comprehensive and effective approach as Contrastive Distance Learning (CDL) to train GAN. This method can be easily extended into different GAN models without any other modification of the backbone.
- We subtly integrate the Siamese modules into the GAN framework with a low computational cost. With its superiority, we alleviate the antagonism between the generator and the discriminator.
- We conduct extensive experiments on three public datasets and demonstrate the versatility of our approach. The results show that CDL can not only address some existing issues in both the generator and discriminator, but also boost the visual quality of the generated images.

2. Preliminaries and Related Works

A GAN is composed of two components: a generator, G , which converts random noise vectors into images, and a discriminator, D , which tries to distinguish between generated and real images. With adversarial training, the generator G is trained to take a latent vector $z \sim P(z)$ and generated target samples $G(z)$ that encourage the capture of the distribution of real images and reduce the discrepancy with the real distribution. The discriminator D as a critic makes a decision score over possible observation sources (either from $G(z)$ or from the empirical data distribution $P_{real}(x)$). Both components have respective loss functions written as follows:

$$L_D = -\mathbb{E}_{x \sim p_{real}} [\log(D(x))] - \mathbb{E}_{z \sim p(z)} [\log(1 - D(G(z)))] \quad (1)$$

$$L_G = -\mathbb{E}_{z \sim p(z)} [\log D(G(z))] \quad (2)$$

The losses defined above originate from the vanilla GAN [1] and are known as the non-saturating constraint. Abundant works have proved that a suitable objective function plays a key role in generation quality and training stability. For example, the hinge loss proposed by Jae Hyun Lim [29], is a very popular redesign loss on GANs and can be written as follows:

$$L_D = -\mathbb{E}_{x \sim p_{real}} [\min(0, D(x) - 1)] - \mathbb{E}_{z \sim p(z)} [\min(0, -D(G(z)) - 1)] \quad (3)$$

$$L_G = -\mathbb{E}_{z \sim p(z)} [D(G(z))] \quad (4)$$

With 1-Lipschitz constrained Wasserstein distance [30], Martin Arjovsky [13] proposes the Wasserstein GAN (W-GAN) to measure distributions, which are fed to the discriminator. Subsequent work has refined this technique in several ways [31,32]. In particular, Takeru Miyato [14] proposed spectral normalization to stabilize the training, which is widely used in many GAN frameworks.

2.1. Regularizations for GANs

Regularization applied in the GANs literature, which encodes some prior knowledge into model training and keeps the predictions consistent, has emerged in recent years. Zhao et al. [21] proposed ICR-GAN and introduced two new techniques, which are abbreviated as bCR and zCR, to improve consistent regularization for GANs. The bCR adds two consistency terms to the discriminator: one is applied in real images, the other is applied in the corresponding sampling from the generator. The zCR augments noise vectors z by a slightly perturbing $\Delta z \sim N(0, \delta_{noise})$ for the generator. Meanwhile, zCR changes the loss function with an additional constraint by maximizing the distance between $G(z)$ and $G(T(z))$ for the discriminator, and motivates the generator to create images with diversity. ICR-GAN improves the quality of generated images indeed compared with CR-GAN [19]; however, it needs some prior knowledge such as image transformations in real data space or noise augmentations in latent space. On the one hand, the discriminator is too sensitive to balance the generator in the bCR and this easily results in training instability and overfitting. On the other hand, the noise enhancement Δz is fixed when fed to the generator and the augmentation cannot make a difference directly to the consistent constraint. It is difficult to guarantee generations with diversity.

To the best of our knowledge, mode seeking regularization (MSR), as presented by Mao et al. [20], has been applied to cGANs for various tasks to alleviate the mode collapse problem. This regularization term encourages generators to generate dissimilar images during training and provide gradients from minor modes to fool the discriminator. MS-GAN can be applied to different conditional image generation tasks for image diversity without sacrificing visual quality. It is a pity that MSR requires labels or extra data as auxiliary information to improve the diversity of synthesized images.

2.2. Characteristic Function Distance for GANs

More recently, characteristic function distance (CFD) [22] reduced to the Integral Probability Metric (IPM) has been reported in the GANs literature and proposes the use of CFD-GAN to improve GANs' performance. Characteristic functions have been widespread in probability theory and successful in two-sample testing [33–35]. The CFD formulates the problem of learning an Implicit Generative Model (IGM) as minimizing the expected distance between characteristic functions. The approximate distance between empirical characteristic functions is seen as a mixture of degenerate distributions with the same weights. The CFD-GAN replaces the Jensen–Shannon (JS) divergence with the CFD and uses the discrepancy expectation between the sampled distribution on real images and generations as an optimizable function. The CFD exhibits desirable mathematical properties such as continuity, differentiability, and weak topology; however, IPM is time-consuming

compared to other distance metrics and CFD-GAN has little improvement in the quality of image synthesis.

2.3. Siamese Network

The Siamese Network proposed by Jane Bromley [36] is a type of metric learning. It is composed of two identical neural networks with sharing weights. The network's map inputs to another space and forms new representations as outputs. In the initial Siamese Network, the loss function is a contrastive loss, which is effective to determine the relationship between paired data. With the rise of deep learning, the Siamese Network has been gradually applied to face detection [37] and object tracking [38,39] in computer vision. In the GANs literature, [40] proposes TraVeLGAN to achieve an image-to-image translation task, which employs the Siamese Network to balance the relationship between the generator and the discriminator by a transformation vector.

Within these optimized approaches mentioned above, GANs has shown its potential for generating natural images, but they are still associated with some problems. Due to mode collapse, samples produced by CFD-GAN often lack diversity and introduce artificial flaws. MS-GAN effectively deals with mode collapse, but it only applies in supervised learning. With augmentations based on the exiting data, ICR-GAN sometimes suffers from unstable training and generates poor-quality images. To alleviate these problems, we introduce MS-GAN into unsupervised learning and propose novel regularizations to the objective function.

3. Methodology: Contrastive Distance Learning

In this section, we present two novel regularizers, Consistent Contrastive Distance (CoCD) and characteristic contrastive distance (ChCD), which are combined and denoted as Contrastive Distance Learning (CDL). To integrate CoCD and ChCD into a GAN architecture perfectly, we utilized the Siamese Network [41] to build the generator and the discriminator. In the training, the Siamese module shares the weights and parameters passing through the model, and we can alleviate the antagonism between the generator and the discriminator by contrastive learning. Our CDL-GAN framework is illustrated in Figure 1. z_1 and z_2 stand for noise vectors. G is composed by two of the same modules S_1 , D is built by two of the same modules S_2 and a Fully Connected (FC) layer. L_{ccd}^G means Characteristic Contrastive Distance (CoCD). S_1 is used to generate two fake images and optimize the L_{ccd}^G . L_{ccd}^D means Characteristic Contrastive Distance (CoCD), the decision layer of S_2 means a projection operation [27] which maps fake or real images into characteristic vectors and gets the metric L_{ccd}^D . S_1 and S_2 both share the weights in the training. The whole framework is used to balance the generative model and discriminative model with CDL.

3.1. Consistent Contrastive Distance

Inspired by ICR-GAN [21] and MS-GAN [20], we focus on the difference between latent consistency regularization (zCR) and mode seeking regularization (MSR). The zCR augments noise vector z to the generator with $T(z)$ by slightly perturbing $\Delta z \sim N(0, \delta_{noise})$, while MSR uses random noise which is variable without controlling parameters. Both regularizations expect to maximize the distance between fake images generated by corresponding noise. To alleviate mode collapse, zCR requires an additional constraint on the discriminator, while MSR just adds a noise coefficient to images' distance. Given the generator with a prior noise input, it is reasonable to explore the effect of the augmentation Δz on the outputs. Considering Δz as a conditional constraint, it is possible for us to make MSR work in unsupervised learning.

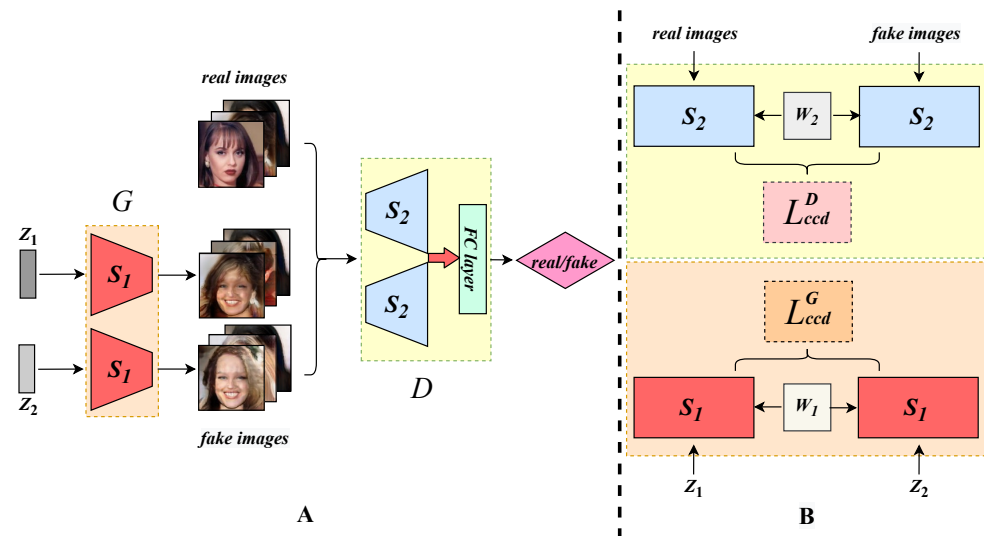


Figure 1. Illustration of the Contrastive Distance Learning-Generative Adversarial Network (CDL-GAN) framework. Part A: the generator, G , consists of a Siamese module, S_1 . The discriminator, D , includes a Siamese module, S_2 , and a fully connected (FC) layer. L_{ccd}^G means Consistent Contrastive Distance (CoCD), while L_{ccd}^D means Characteristic Contrastive Distance (ChCD). S_1 shares the weights W_1 when noise vector z_1 and z_2 are fed to G . S_2 shares the weights W_2 when real and fake images pass through D .

To integrate MSR into unconditional GANs, we propose Consistent Contrastive Distance (CoCD). We augment noise z to the generator by setting a hyperparameter γ and yield $A(z) = (1 + \gamma) * z, \gamma \in (0, 1)$. To this end, $\Delta z = \gamma * z$. What we need to emphasize here is that we only augment the amplitude of noise vector z to keep them consistent. In the training processing, z and $A(z)$ learn feature distribution from the same image. When the augment Δz is small enough, we expect that the distance between $G(z)$ and $G(A(z))$ is large enough to encourage generations to be discrepant. Taking Δz into consideration, the distance metric can be written as follows:

$$L_{ccd}^G = \arg \max_G \left(\frac{L_2(G(A(z))) - L_2(G(z))}{L_2(\Delta z)} \right) \quad (5)$$

where L_2 means the second-norm. With the adversarial mechanism, consistent contrastive distance as a regularization term can be appended into the original objective function. Here, we take Equation (4) for example and the generator's loss can be written as follows:

$$L'_G = L_G - \lambda_{gen} L_{ccd}^G \quad (6)$$

where λ_{gen} denotes the control of the weights and highlights the importance of the regularizer. In the training process, Δz is controlled by γ after a noise vector is generated randomly. We can modulate the parameter γ to achieve the effectiveness of CoCD.

3.2. Characteristic Contrastive Distance

To stabilize training, we utilize Characteristic Function Distance (CFD) introduced in CFD-GAN [22] as a regularization for the discriminator. Meanwhile, the transformation vector introduced in TraVeL-GAN [40], which uses a function to represent high-level semantics in some latent space, is aimed at mapping images to some space with the same relationship between the original and generated versions. TraVeL-GAN gives us a feasible scheme to add the CFD regularization to the discriminator. We extract semantic information by a characteristic function and utilize a transformation vector to achieve Characteristic Contrastive Distance (ChCD). To elaborate, we firstly extract the output of the second-to-last layer as high-level semantic information when a real image or generated

image is encoded by the discriminator. Then, we take a characteristic function as the transformation mode and utilize the transformation vector to learn characteristic semantic information. Next, we turn characteristic semantic information into characteristic vectors with a finite-dimensional approximation in Euler space. Finally, we compute the distance between real characteristic vectors and generated characteristic vectors.

Unlike the density function, the Euler space always exists with uniform continuity, differentiability, and boundedness. From the perspective of manifold learning, a characteristic vector is regarded as an essential element of images, and we could get the discrepancy between images by contrasting the distance between characteristic vectors. For GAN, ChCD can keep the parameters of the discriminator continuous and differentiable almost everywhere and provide a more informative signal to the discriminator for feature representations. The proofs of CFD properties are stated in CFD-GAN [21], which the reader is highly encouraged to read.

Letting t be the input argument of the characteristic vector $V : \{v_1, \dots, v_n\}$, the characteristic function $\hat{\psi}$ is a weighted sum of characteristic vectors transformed in Euler space:

$$\hat{\psi}(t) = \frac{1}{n} \sum_{j=1}^n e^{i(t \odot v_j)} \quad (7)$$

where $i = \sqrt{-1}$, $|e^{ix}| \leq 1$, $t \in R^d$, $\odot$ means vector dot product, and t is a random variable of a degenerated sampling distribution δ_V . Given $X := \{x_1, \dots, x_n\}$ and $Y := \{y_1, \dots, y_n\}$ with $x_i, y_i \in R^d$ are samples from the distributions $\mathbb{P}$ and $\mathbb{Q}$, respectively, and let $t_1, \dots, t_k$ be samples from δ_V . We define the characteristic contrastive distance between $V_{\mathbb{P}}$ and $V_{\mathbb{Q}}$ as

$$L_{ccd}^D(V_{\mathbb{P}}, V_{\mathbb{Q}}) = \frac{1}{k} \sum_{i=1}^k \|\hat{\psi}_{\mathbb{P}}(t_i) - \hat{\psi}_{\mathbb{Q}}(t_i)\|^2 \quad (8)$$

where $\hat{\psi}_{\mathbb{P}}$ and $\hat{\psi}_{\mathbb{Q}}$ are the characteristic functions of characteristic vectors, computed by X and Y , respectively. With the characteristic contrastive distance, the new objective function of the discriminator can be written as follows:

$$L'_D = L_D + \lambda_{dis} L_{ccd}^D \quad (9)$$

where λ_{dis} denotes the control of the importance of the L_{ccd}^D and L_D denotes the original loss.

3.3. Enhancement with the Siamese Modules

Combining Contrastive Distance Learning (CDL) with the Siamese modules, we can make use of the same structure to share the weights and handle the relationship between paired data effectively. The method is shown in greater detail in Algorithm 1. On the one hand, Consistent Contrastive Distance (CoCD) offers a virtuous cycle for the generator exploiting more modes. On the other hand, Characteristic Contrastive Distance (ChCD) forces the discriminator to focus on meaningful visual information. The goal of the Siamese module, S_1 , is cooperative with the generator, G . Meanwhile, the discriminator, D , is constrained by the module S_2 , to make training more bidirectionally stable. In the training process, the generator creates two parts of fake data, while the discriminator needs to consider the difference of each real image with the two corresponding generated images. CDL can be balanced by adjusting the parameters of λ_{gen} and λ_{dis} .

Algorithm 1 Contrastive Distance Learning (CDL)

Input: parameters of generator θ_G and discriminator θ_D , consistent contrastive distance coefficient λ_{gen} and characteristic contrastive distance coefficient λ_{dis} , noise z and noise augment parameter γ .

```

1: for number of training iterations do
2:   for  $t = 1$  to  $N_D$  do
3:     sample batch  $Z \sim P(z)$ ,  $X \sim P_{real}(x)$ 
4:     sample augment noise  $\Delta z = \gamma * z$ 
5:     augment latent vector  $A(z) = z + \Delta z$ 
6:      $L_D \leftarrow (G(z)) - D(x) + D(G(A(z))) - D(x)$ 
7:      $L_{dis} \leftarrow L_{ccd}^D(V_{G(z)}, V_x) + L_{ccd}^D(V_{G(A(z))}, V_x)$ 
8:      $\theta_D \leftarrow \text{AdamOptimizer}(L_D + \lambda_{dis}L_{dis})$ 
9:   end for
10:  sample batch  $Z \sim P(z)$ 
11:  sample augment noise  $\Delta z = \gamma * z$ 
12:  augment latent vector  $A(z) = z + \Delta z$ 
13:   $L_G \leftarrow -D(G(z)) - D(G(A(z)))$ 
14:   $L_{gen} \leftarrow -L_{ccd}^G$ 
15:   $\theta_G \leftarrow \text{AdamOptimizer}(L_G - \lambda_{gen}L_{gen})$ 
16: end for

```

4. Experiments

To validate our proposed CDL-GAN method, we conducted extensive quantitative and qualitative experiments to evaluate different aspects. First, we compare CDL-GAN to several existing optimized works, ICR-GAN [21], CFD-GAN [22], and WGAN-GP [31], with the same GAN backbone for three public datasets. We highlight here that CDL-GAN is motivated by ICR-GAN and CFD-GAN, while WGAN-GP is effective for stabilizing GAN training with 1-Lipschitz; therefore, it is necessary for us to conduct comparisons with them. Then, we applied CoCD, ChCD, and CDL, respectively, to a recent state-of-the-art baseline SNGAN [14] with two different hyperparameter sets of the Adam optimizer [42]. Next, we re-implemented SNGAN with our approach to analyze the training time for different datasets. Finally, we conducted studies based on DCGAN [43] and ICR-GAN to evaluate CDL-GAN's mode recovery ability. For fairness, we emphasize that all GAN models as discussed are unconditional and our operational procedures were under the same training conditions with a uniform code base. In the same experiment, we chose the same GAN backbone. We reproduced the existing models using the description in the corresponding works.

4.1. Settings and Evaluation Metrics

We evaluate our models against the above existing models for three public datasets: MNIST [44], CIFAR-10 [45], and CelebA [46]. For preprocessing of data sets, we follow the detailed settings in [47]. MNIST contains 70 K 28×28 handwritten digits with 10 labels; 60 K for training and 10 K for testing. We use all unlabeled training images in our experiments, resized to 32×32 . CIFAR-10 consists of 60 K 32×32 natural images in 10 classes, out of which 50 K for training and 10 K for testing. We use all the training images without labels. For CelebA, we use the aligned face version in two resolutions: about 200 K images at a resolution of 64×64 and 30K images reshaped to size 128×128 . In this paper, we regard the higher resolution images of CelebA as CelebA-HD.

To assess the quality of generated images against corresponding real images, we adopted the Fréchet Inception Distance (FID) [28] as a standard metric, which is used to measure the distance between generated and real image features and proved to correlate well with human evaluation. In our experiments, we calculated FID scores on different datasets with different image numbers: we used 10 K generated images vs. 10 K real images on CIFAR-10, 50 K vs. 50 K face images on CelebA in the size of 64×64 , and 3 K

vs. 3 K on the high-resolution CelebA. Lower FID scores indicate better quality of the synthetic images.

For all experiments, we use a single Tesla V100 GPU with our implementation in PyTorch. Meanwhile, we chose the Adam optimizer [42] with the learning rate of 0.0002 and set the batch size of images as 64. For iterations, the number of discriminator steps was 5 per generator step. For hyperparameters, we initialized $\gamma = 0.05$, $\lambda_{dis} = 10$, and $\lambda_{gen} = 10$. For objective function, we used hinge loss as a basic criterion except for CFD-GAN, because CFD-GAN utilizes a characteristic function distance as its objective function.

To integrate CDL into GAN architecture perfectly, we updated the DCGAN and SNGAN backbone with Siamese modules, allowing the generator to transform the global structure of an input image by ChCD regularization. We further enhanced the training stability and applied ChCD regularization and switched the depth-wise concatenation to adaptive instance normalization in the discriminator. Our generator consists of four down-sampling blocks, four intermediate blocks, and four upsampling blocks, all of which inherit preactivation residual units. Our discriminator is a projection discriminator [27], which contains multiple linear output branches. The discriminator contains six pre-activation residual blocks with leaky ReLU.

4.2. Results

For improved image synthesis, we achieved a coincident and remarkable improvement in FID scores across three public datasets over the existing models mentioned above. All the models use the DCGAN as the backbone. As seen in Table 1, CDL-GAN improves FID scores on CIFAR-10 by 1.54 points over ICR-GAN, 18.75 points over CFD-GAN, and 7.96 points over WGAN-GP. On CelebA dataset, CDL-GAN achieved an improvement of 2.25 points over ICR-GAN, 7.30 points over CFD-GAN, and 4.67 points over WGAN-GP at a resolution of 64×64 . On the high-resolution of CelebA, CDL-GAN improved by 3.03 points over ICR-GAN, 9.20 points over CFD-GAN, and 6.55 points over WGAN-GP. The results in Figure 2 show that CDL-GAN significantly outperforms all other models in terms of FID scores. We determined the FID scores in each dataset with five validations for different models. The improvement is still significant compared to the measurement variance. The height difference of each rectangle suggests that our method is more reliable than the existing models in terms of the quality of image synthesis.

Table 1. Fréchet Inception Distance (FID) scores for CDL-GAN and the existing models.

Models	MNIST (32*32)	CIFAR-10 (32*32)	CelebA (64*64)	CelebA-HD (128*128)
ICR-GAN	0.55	15.87	15.43	19.56
CFD-GAN	0.53	33.08	20.48	25.73
WGAN-GP	0.69	22.29	17.85	23.98
CDL-GAN (ours)	0.48	14.33	13.18	16.53

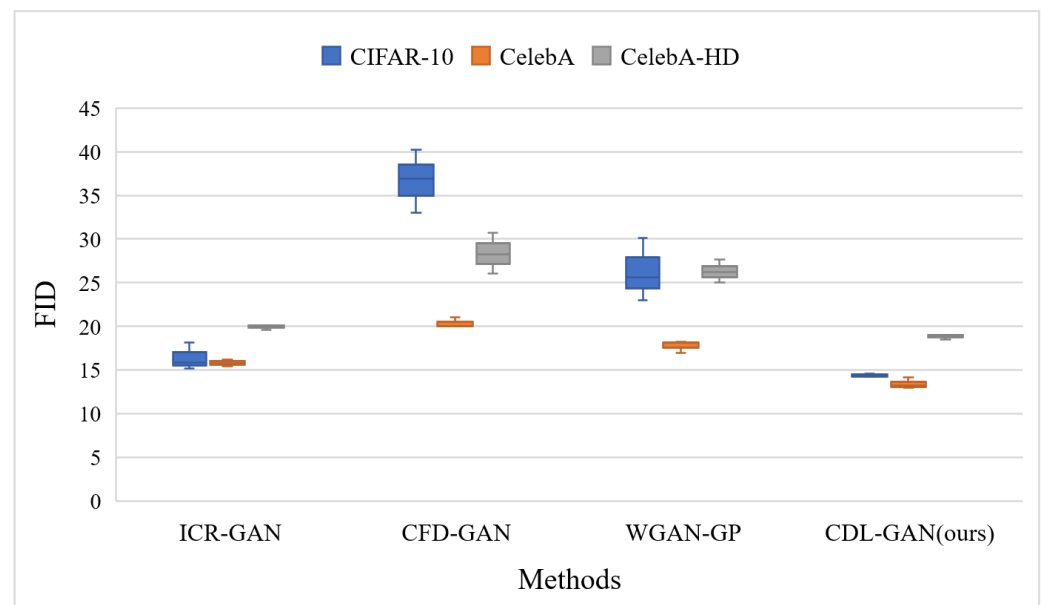


Figure 2. FID scores (lower is better) for CDL-GAN and the existing models for different datasets. The FID scores in each dataset with 5 validations for different models.

For improved training stability, we verified it in two aspects: using the Adam parameters (β_1, β_2) to evaluate the sensitivity of model performance based on the SNGAN [14] backbone and observing the convergence rate of GAN training without any hyperparameter tuning. As seen in Table 2, in comparison to the baseline SNGAN, SNGAN + CDL achieves lower FID scores than SNGAN. When the Adam optimizer hyperparameter set was (0.0, 0.9), SNGAN + CDL improved by 2.89 points over SNGAN. Meanwhile, SNGAN + CDL achieved an improvement of 4.31 points over SNGAN with the Adam parameters (0.5, 0.999). Furthermore, the variability in FID scores for SNGAN + CDL on different Adam parameters was 0.49 points, which is smaller than SNGAN with 1.97 points. In Figure 3, we show the relationship between FID scores and training iterations on the baseline SNGAN. It is obvious that SNGAN + CDL converges faster in the early training period and achieves better FID scores in the end. Thereby, our method possesses better robustness in all of these settings.

Table 2. FID scores (lower is better) among SNGAN, SNGAN + CoCD, SNGAN + ChCD, and SNGAN + CDL on CelebA-HD. “SNGAN + CoCD” means combining SNGAN with the consistent contrastive distance, “SNGAN + ChCD” means combining SNGAN with the characteristic contrastive distance, and “SNGAN + CDL” means combining SNGAN with contrastive distance learning.

(β_1, β_2)	SNGAN	SNGAN + CoCD	SNGAN + ChCD	SNGAN + CDL
(0.0, 0.9)	12.97	12.12	11.35	10.08
(0.5, 0.999)	14.88	13.33	12.98	10.27

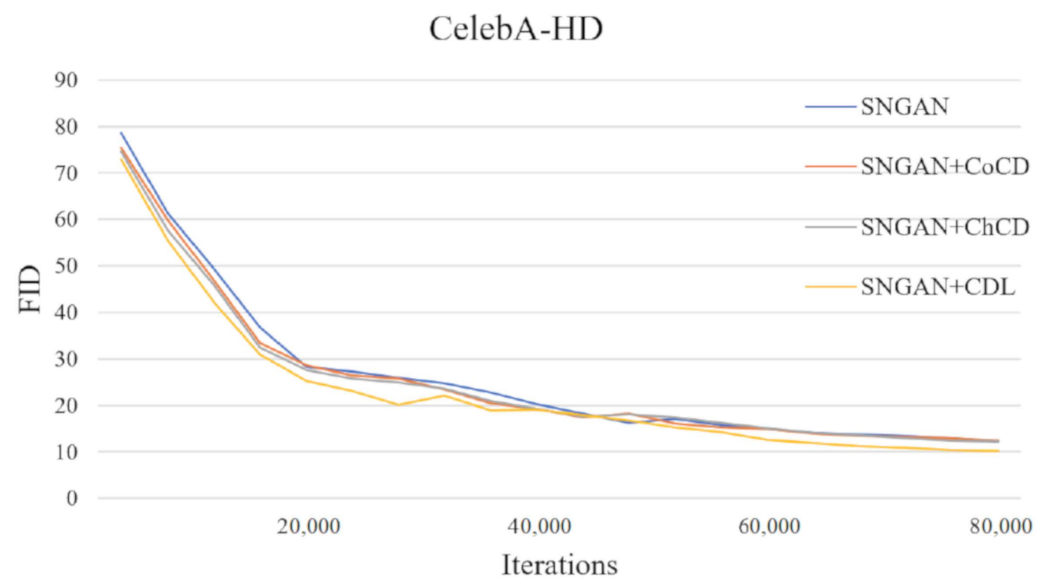


Figure 3. Training stability among SNGAN, SNGAN + CoCD, SNGAN + ChCD, and SNGAN + CDL on CelebA-HD with the hyperparameter set of (0.0, 0.9). All the models utilize the SNGAN backbone with residual blocks.

For a lower computational cost, we profile CDL-GAN’s training time with SNGAN backbone for 100 generator update steps, which is similar to [14]. From Figure 4, we can see that our approach takes up minimal time at less than 0.1% of training time per update for all datasets. GANs training with less time taken means decreasing computational costs, and we can easily integrate CDL into the existing GAN frameworks without worrying about extra burden.

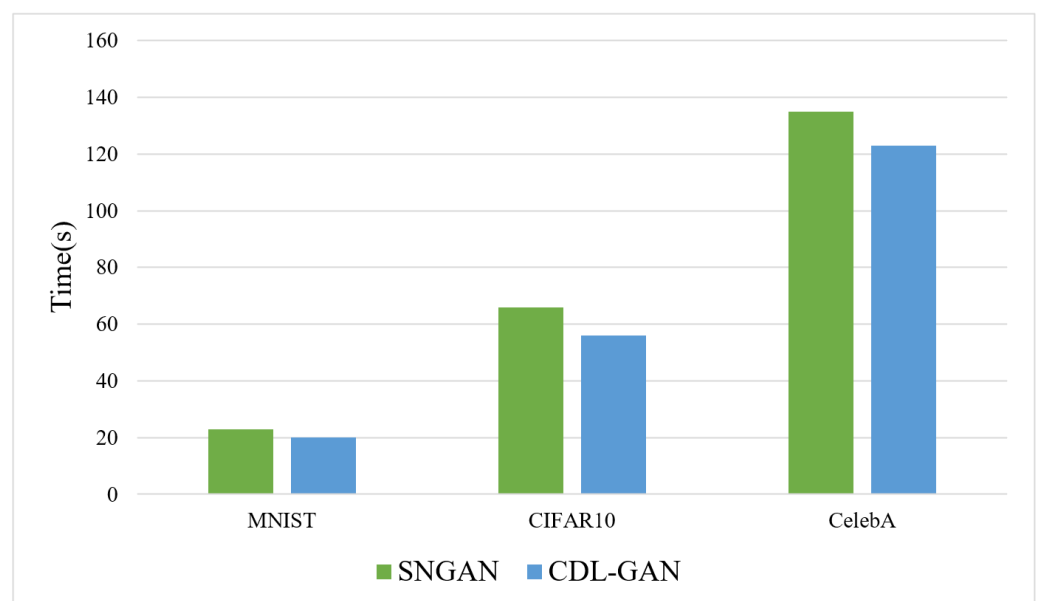


Figure 4. Training times for 100 generator update steps on different datasets. CDL-GAN and SNGAN utilize the same resblocks to build the backbone.

For mitigating mode collapse, we followed the procedure in [43,48] evaluated mode recovery with DCGAN [43] and ICR-GAN [21]. We utilized a pre-trained MNIST classifier and the Stacked MNIST [48] dataset with 1000 possible modes to conduct a comparative analysis. As seen in Table 3, K indicates the size of the discriminator relative to the generator, $D_{KL}(p||q)$ means the KL divergence between the generated mode distribution p

and optimal uniform distribution of the mode q . The results show that CDL-GAN recovers more modes for all K and has a lower KL divergence with the ideal uniform distribution than both DCGAN and ICR-GAN. More modes recovery from CDL-GAN means that our method possesses better generation diversity.

To support our improvements directly in the observation, we also provide image samples randomly generated by CDL-GAN and the existing models for all datasets in Figures 5–8. It can be seen from these results that the generations yielded by CDL-GAN are more explicit and more authentic.

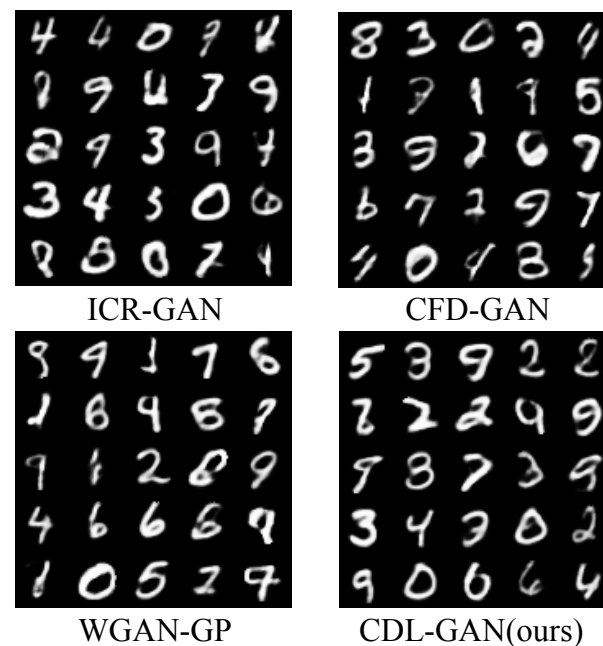


Figure 5. Image samples from different models on MNIST dataset.



Figure 6. Image samples from different models on CIFAR-10 dataset.

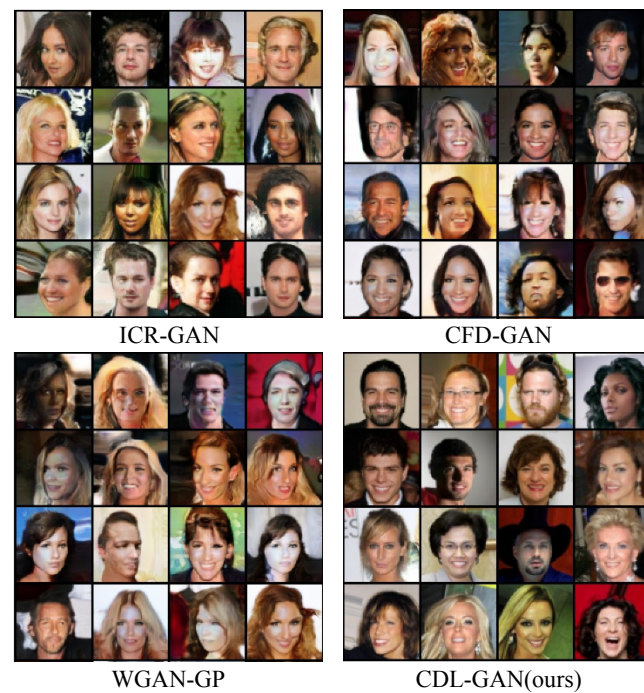


Figure 7. Image samples from different models on CelebA dataset.

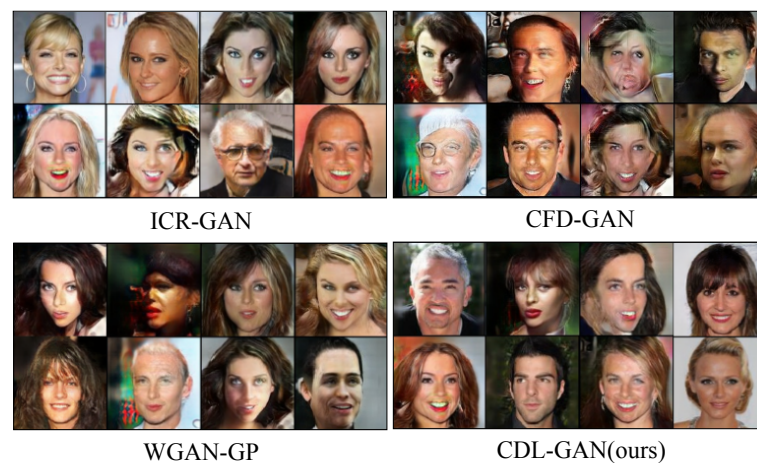


Figure 8. Image samples from different models on CelebA-HD dataset.

Table 3. Number of modes (higher is better) recovered by the generator on the Stacked MNIST dataset.

Metric	K	DCGAN	ICR-GAN	CDL-GAN
Modes	1/4	30.64	48.47	50.50
Modes	1/2	605.17	703.25	725.88
$D_{KL}(p q)$	1/4	5.48	4.96	4.55
$D_{KL}(p q)$	1/2	1.97	1.64	1.45

4.3. Ablation Study

To verify the effectiveness of our proposed CoCD and ChCD, we further analyzed the training stability results. As seen in Table 2, we also added CoCD and ChCD into SNGAN [14], respectively, and the effect is cogent. When the Adam parameter set was (0.0, 0.9), SNGAN + CoCD improved FID scores by 0.85 points over SNGAN, SNGAN + ChCD improved by 1.62 points over SNGAN. Meanwhile, SNGAN + CoCD improved by

1.55 points over SNGAN with the Adam parameters (0.5, 0.999), while an improvement of 2.60 was found for SNGAN + ChCD. With two different ranges of the Adam parameter, the FID score discrepancies were 1.21 points for SNGAN + CoCD, 0.93 points for SNGAN + ChCD, which are lower than SNGAN with 1.97 points. The results in Figure 3 also present a faster convergence in the early training period for CoCD and ChCD applied to SNGAN. The best FID scores achieved by SNGAN + CDL suggest that CoCD and ChCD can cooperate with each other to improve GANs' performance. From Figure 9, we can see the visualization results of the generated samples. When CoCD and ChCD are respectively attached to the SNGAN model, the generated images have better quality. In particular, when CDL is attached to SNGAN, the generated face images are more authentic and recognizable.

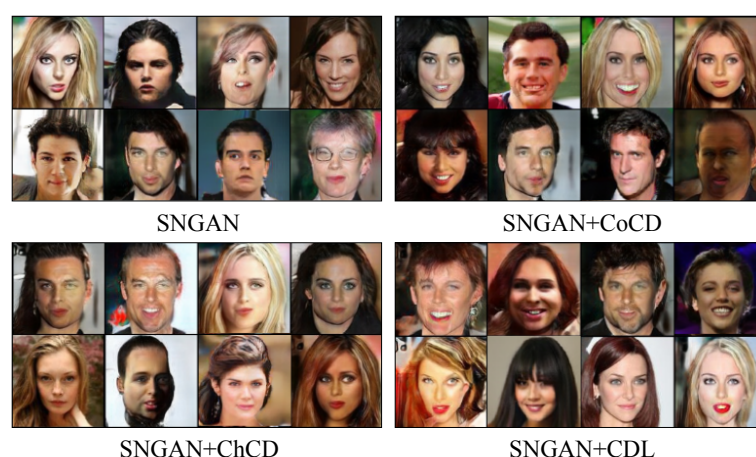


Figure 9. Image samples from SNGAN, SNGAN + CoCD, SNGAN + ChCD, and SNGAN + CDL on CelebA-HD.

5. Conclusions and Discussion

In this work, we presented Contrastive Distance Learning (CDL) as a novel optimized approach for GANs. For clarity, CDL includes two regularizations: Consistent Contrastive Distance (CoCD) and Characteristic Contrastive Distance (ChCD). Furthermore, we added the Siamese modules to the GAN backbone to balance the relationship between the generator and discriminator.

Extensive experiments have shown that our approach is practical and versatile. With the DCGAN backbone, CDL-GAN not only achieved lower FID scores than the existing optimized methods, but also effectively alleviated the mode collapse problem. With the SNGAN backbone, CDL-GAN also achieved better FID scores and improved training stability compared to SNGAN. Meanwhile, all the experiments show that CDL can be integrated into different GAN backbones, such as DCGAN and SNGAN, which indicates that CDL has universality in the GANs literature. All the results prove that CDL leads to significant improvements in the image synthesis task and provides an effective and alternative means for GAN training.

As for future work, we hope to explore CDL for other image tasks such as image-to-image translation and try to integrate mutual information into ChCD to improve GAN performance.

Author Contributions: Conceptualization, Y.Z. (Yingbo Zhou) and P.Z.; methodology, Y.Z. (Yingbo Zhou) and W.T.; software, Y.Z. (Yingbo Zhou) and P.Z.; validation, Y.Z. (Yongxin Zhu) and W.T.; formal analysis, Y.Z. (Yingbo Zhou) and W.T.; investigation, Y.Z. (Yingbo Zhou) and P.Z.; resources, Y.Z. (Yongxin Zhu); data curation, Y.Z. (Yingbo Zhou) and P.Z.; writing—original draft preparation, Y.Z. (Yingbo Zhou); writing—review and editing, Y.Z. (Yongxin Zhu); visualization, Y.Z. (Yingbo Zhou); supervision, W.T.; project administration, Y.Z. (Yingbo Zhou); funding acquisition, Y.Z. (Yongxin Zhu). All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported by the National Key Research and Development Project (Grant No.2019YFC0117302), the National Natural Science Foundation of China (Grant No. 61772331), National Key Research and Development Program of China (Grant No. 2018YFA0701500), Natural Science Foundation of China (grant No.U2032125), Natural Science Foundation of China (grant No.U1831118), Shanghai Municipal Science and Technology Commission (grant No. 19511131202), Independent Deployment Project of Shanghai Advanced Research Institute (Grant E0560W1ZZ0, E052891ZZ1) and also supported in part by grant PKX2019-D02 from Pudong Industry-University-Research Project.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

- Goodfellow, I.J.; Pouget-Abadie, J.; Mirza, M.; Bing, X.; Bengio, Y. Generative adversarial nets. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 8–13 December 2014; pp. 2672–2680.
- Mirza, M.; Osindero, S. Conditional generative adversarial nets. *arXiv* **2014**, arXiv:1411.1784.
- Huang, X.; Li, Y.; Poursaeed, O.; Hopcroft, J.; Belongie, S. Stacked generative adversarial networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 5077–5086.
- Odena, A.; Olah, C.; Shlens, J. Conditional image synthesis with auxiliary classifier gans. In Proceedings of the International Conference on Machine Learning, Sydney, Australia, 6–11 August 2017; pp. 2642–2651.
- Wang, Y.; Zhang, L.; Van De Weijer, J. Ensembles of generative adversarial networks. *arXiv* **2016**, arXiv:1612.00991.
- Hoang, Q.; Nguyen, T.D.; Le, T.; Phung, D. Multi-generator generative adversarial nets. *arXiv* **2017**, arXiv:1708.02556.
- Nguyen, T.; Le, T.; Vu, H.; Phung, D. Dual discriminator generative adversarial nets. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 2670–2680.
- Larsen, A.B.L.; Sønderby, S.K.; Larochelle, H.; Winther, O. Autoencoding beyond pixels using a learned similarity metric. In Proceedings of the International Conference on Machine Learning, New York, NY, USA, 19–24 June 2016; pp. 1558–1566.
- Makhzani, A.; Shlens, J.; Jaitly, N.; Goodfellow, I.; Frey, B. Adversarial autoencoders. *arXiv* **2015**, arXiv:1511.05644.
- Dumoulin, V.; Belghazi, I.; Poole, B.; Mastropietro, O.; Lamb, A.; Arjovsky, M.; Courville, A. Adversarially learned inference. *arXiv* **2016**, arXiv:1606.00704.
- Wang, X.; Wang, X. Unsupervised Domain Adaptation with Coupled Generative Adversarial Autoencoders. *Appl. Sci.* **2018**, *8*, 2529. [[CrossRef](#)]
- Kwak, J.g.; Ko, H. Unsupervised Generation and Synthesis of Facial Images via an Auto-Encoder-Based Deep Generative Adversarial Network. *Appl. Sci.* **2020**, *10*, 1995. [[CrossRef](#)]
- Arjovsky, M.; Chintala, S.; Bottou, L. Wasserstein gan. *arXiv* **2017**, arXiv:1701.07875.
- Miyato, T.; Kataoka, T.; Koyama, M.; Yoshida, Y. Spectral normalization for generative adversarial networks. *arXiv* **2018**, arXiv:1802.05957.
- Salimans, T.; Zhang, H.; Radford, A.; Metaxas, D. Improving GANs using optimal transport. *arXiv* **2018**, arXiv:1803.05573.
- Li, C.L.; Chang, W.C.; Cheng, Y.; Yang, Y.; Póczos, B. Mmd gan: Towards deeper understanding of moment matching network. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 2203–2213.
- DeVries, T.; Taylor, G.W. Improved regularization of convolutional neural networks with cutout. *arXiv* **2017**, arXiv:1708.04552.
- Wei, X.; Gong, B.; Liu, Z.; Lu, W.; Wang, L. Improving the improved training of wasserstein gans: A consistency term and its dual effect. *arXiv* **2018**, arXiv:1803.01541.
- Odena, A.; Zhang, H.; Lee, H.; Zhang, Z. Consistency Regularization for Generative Adversarial Networks. *arXiv* **2019**, arXiv:1910.12027.
- Mao, Q.; Lee, H.Y.; Tseng, H.Y.; Ma, S.; Yang, M.H. Mode seeking generative adversarial networks for diverse image synthesis. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 1429–1437.
- Zhao, Z.; Singh, S.; Lee, H.; Zhang, Z.; Odena, A.; Zhang, H. Improved consistency regularization for gans. *arXiv* **2020**, arXiv:2002.04724.
- Ansari, A.F.; Scarlett, J.; Soh, H. A Characteristic Function Approach to Deep Implicit Generative Modeling. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 7478–7487.
- Mescheder, L.; Nowozin, S.; Geiger, A. The numerics of gans. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 1825–1835.
- Daskalakis, C.; Ilyas, A.; Syrgkanis, V.; Zeng, H. Training gans with optimism. *arXiv* **2017**, arXiv:1711.00141.

25. Prasad, H.; LA, P.; Bhatnagar, S. Two-timescale algorithms for learning Nash equilibria in general-sum stochastic games. In Proceedings of the 2015 International Conference on Autonomous Agents and Multiagent Systems, Istanbul, Turkey, 4–8 May 2015; pp. 1371–1379.
26. Yadav, A.; Shah, S.; Xu, Z.; Jacobs, D.; Goldstein, T. Stabilizing adversarial nets with prediction methods. *arXiv* **2017**, arXiv:1705.07364.
27. Miyato, T.; Koyama, M. cGANs with projection discriminator. *arXiv* **2018**, arXiv:1802.05637.
28. Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; Hochreiter, S. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 6626–6637.
29. Lim, J.H.; Ye, J.C. Geometric gan. *arXiv* **2017**, arXiv:1705.02894.
30. Villani, C. *Optimal Transport: Old and New*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2008; Volume 338.
31. Gulrajani, I.; Ahmed, F.; Arjovsky, M.; Dumoulin, V.; Courville, A.C. Improved training of wasserstein gans. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 5767–5777.
32. Miyato, T.; Maeda, S.i.; Koyama, M.; Ishii, S. Virtual adversarial training: A regularization method for supervised and semi-supervised learning. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *41*, 1979–1993. [[CrossRef](#)]
33. Chwialkowski, K.P.; Ramdas, A.; Sejdinovic, D.; Gretton, A. Fast two-sample testing with analytic representations of probability measures. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 7–12 December 2015; pp. 1981–1989.
34. Epps, T.; Singleton, K.J. An omnibus test for the two-sample problem using the empirical characteristic function. *J. Stat. Comput. Simul.* **1986**, *26*, 177–203. [[CrossRef](#)]
35. Heathcote, C. A test of goodness of fit for symmetric random variables1. *Aust. J. Stat.* **1972**, *14*, 172–181. [[CrossRef](#)]
36. Bromley, J.; Guyon, I.; Lecun, Y.; Sckinger, E.; Shah, R. Signature Verification Using a Siamese Time Delay Neural Network. In Proceedings of the Advances in Neural Information Processing Systems 6, 7th NIPS Conference, Denver, CO, USA, 22–23 November 1993.
37. Nair, V.; Hinton, G.E. Rectified linear units improve restricted boltzmann machines. In Proceedings of the 27th International Conference on Machine Learning, Haifa, Israel, 21–24 June 2010.
38. Bertinetto, L.; Valmadre, J.; Henriques, J.F.; Vedaldi, A.; Torr, P.H.S. Fully-Convolutional Siamese Networks for Object Tracking. In Proceedings of the Computer Vision—ECCV 2016 Workshops, Amsterdam, The Netherlands, 8–10. 15–16 October 2016; pp. 850–865.
39. Xu, Z.; Luo, H.; Hui, B.; Chang, Z.; Ju, M. Siamese Tracking with Adaptive Template-Updating Strategy. *Appl. Sci.* **2019**, *9*, 3725 [[CrossRef](#)]
40. Amodio, M.; Krishnaswamy, S. Travelgan: Image-to-image translation by transformation vector learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 8983–8992.
41. Zagoruyko, S.; Komodakis, N. Learning to compare image patches via convolutional neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 4353–4361.
42. Da, K. A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
43. Radford, A.; Metz, L.; Chintala, S. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv* **2015**, arXiv:1511.06434.
44. LeCun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. Gradient-based learning applied to document recognition. *Proc. IEEE* **1998**, *86*, 2278–2324. [[CrossRef](#)]
45. Krizhevsky, A.; Hinton, G. Learning multiple layers of features from tiny images. In *Handbook of Systemic Autoimmune Diseases*; Elsevier: Amsterdam, The Netherlands, 2009.
46. Liu, Z.; Luo, P.; Wang, X.; Tang, X. Deep Learning Face Attributes in the Wild. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015.
47. Kurach, K.; Lucic, M.; Zhai, X.; Michalski, M.; Gelly, S. A Large-Scale Study on Regularization and Normalization in GANs. *arXiv* **2019**, arXiv:1807.04720.
48. Metz, L.; Poole, B.; Pfau, D.; Sohl-Dickstein, J. Unrolled Generative Adversarial Networks. *arXiv* **2016**, arXiv:1611.02163.