Neural Characteristic Function Learning for Conditional Image Generation

Shengxi Li¹, Jialu Zhang¹, Yifei Li¹, Mai Xu*¹, Xin Deng², Li Li¹

¹School of Electronic and Information Engineering, Beihang University, Beijing, China
²School of Cyber Science and Technology, Beihang University, Beijing, China
{LiShengxi, JialuZhang, leafy, MaiXu, cindydeng, lili2005}@buaa.edu.cn

Abstract

The emergence of conditional generative adversarial networks (cGANs) has revolutionised the way we approach and control the generation, by means of adversarially learning joint distributions of data and auxiliary information. Despite the success, cGANs have been consistently put under scrutiny due to their ill-posed discrepancy measure between distributions, leading to mode collapse and instability problems in training. To address this issue, we propose a novel conditional characteristic function generative adversarial network (CCF-GAN) to reduce the discrepancy by the characteristic functions (CFs), which is able to learn accurate distance measure of joint distributions under theoretical soundness. More specifically, the difference between CFs is first proved to be complete and optimisation-friendly, for measuring the discrepancy of two joint distributions. To relieve the problem of curse of dimensionality in calculating CF difference, we propose to employ the neural network, namely neural CF (NCF), to efficiently minimise an upper bound of the difference. Based on the NCF, we establish the CCF-GAN framework to explicitly decompose CFs of joint distributions, which allows for learning the data distribution and auxiliary information with classified importance. The experimental results on synthetic and real-world datasets verify the superior performances of our CCF-GAN, on both the generation quality and stability.

1. Introduction

Generative adversarial network (GAN) has been the workhorse in deep generative models since its birth for image generation [16], and its popularity arises from the capability of generating clear and realistic images from merely small dimensions. Despite success, the original architecture of GAN only allows for randomly generating images from Gaussian noise, and an important variant of GANs aims to control the generation by pre-defined auxiliary information (e.g., the class labels or texts), constituting the conditional GAN (cGAN). Taking advantages of the auxiliary information, cGANs have been proved to be capable of enhancing the realistic image generation that is conditioned on extra semantic cues [42, 32, 33]. Therefore, the past few years have witnessed the extensive applications of cGANs, including class-conditioned generation [31, 37], style transfer [55], text-to-image translation [42, 51], to name but a few.

Generally speaking, cGANs establish a joint distribution between data $\mathcal{X}$ and auxiliary information $\mathcal{Y}$, i.e., $\{\mathcal{X}, \mathcal{Y}\} \sim p(x, y)$. Most cGANs agreed on the design of the generator network, in which the auxiliary information is embedded to the input noise [31] or the inter-mediate layers of the generator [11, 37, 9, 50, 40, 4, 33]. As such, the generator aims to sample from the joint distribution $p(x, y)$. On the other hand, for designing the discriminator, the way we formulate the conditional distribution tells the existing cGANs apart, because $p(x, y)$ can be formulated by either $p(x|y)p(y)$ or $p(y|x)p(x)$. The former calls for transforming the auxiliary information $\mathcal{Y}$ into the discriminator so as to predict $p(x|y)$, and this can be achieved by concatenating with $\mathcal{X}$ as input [31, 10, 44], or embedding $\mathcal{Y}$ to hidden layers of the discriminator [42, 51]. The latter, however, requires the discriminator to predict the auxiliary information $p(y|x)$, by for example, additional explicit classifiers [37, 15, 21] or implicit projections [33, 50, 32, 4]. Despite being able to control the generation by pre-defined auxiliary cues, applying cGANs in practice has been significantly restricted owing to their mode collapse [48, 52, 23] and instability [37, 33] problems in training, thus impeding the consistent improvement in the realistic image generation.

Indeed, most discriminators of cGANs build upon the cross-entropy adversarial loss, with an equivalence to the Jensen-Shannon (JS) divergence between generated and real data distributions [2]. Unfortunately, it has been verified both theoretically and empirically that the JS divergence,
which compares two distributions in a “bin-to-bin” manner [25], can easily max out when the two distributions are mis-aligned or supported by low dimensions [2]. Consequently, there exists an issue of gradient vanishing in the discriminator, which misleads the generator to simply learn fixed patterns or completely break down in training [2, 3]. For unconditional generation, this issue has been elegantly addressed by introducing a broad class of distance metrics called integral probability metric (IPM) [35]. Under the umbrella of the theoretical completeness of IPMs, the discriminator operates as certain bounded functions to compare distributions in a “cross-bin” style [25], such that smooth and sufficient gradient can be provided for unconditional generation.

Therefore, it is intuitive to apply IPMs to conditional generation, benefiting from the theoretical completeness of IPMs to stably and consistently improve the generation. However, it is non-trivial to design an IPM-cGAN, due to the non-linear coupling between the data and auxiliary information. In other words, the bounded function of the discriminator prohibits explicitly modelling $p(x|y)$ or $p(y|x)$ for conditional generation. Several attempts were proposed to concatenate $X$ and $Y$ as an augmented random variable $X'$, and equivalently train the cGAN by an unconditional IPM-GAN [54]. However, it is problematic to straightforwardly combine two random variables at different semantic levels, whereby its deficiency has been proved in many cGANs [28, 33]. Although several cGANs employed certain IPMs, e.g., the Wasserstain distance in their implementations [34, 44, 33], their very basic theories were established upon the cross-entropy form (equivalent to the JS divergence), thus still suffering from the mode collapse and instability problems caused by the “bin-to-bin” comparison. More importantly, the above cGANs are established upon the existence of probability density functions (pdfs) of random variables. This premise, oftentimes taken for granted without verification, may not hold in practice, especially when real-world data such as images and videos essentially reside on low-dimensional manifolds [24, 36].

In this paper, we propose a novel cGAN architecture upon the characteristic function (CF) of random variables, i.e., conditional characteristic function GAN (CCF-GAN). We also noticed several works [1, 30] built upon the CF to achieve enhanced unconditional generation. Those methods, however, by first embedding the data distributions into latent spaces, are problematic in learning joint distributions of the data and auxiliary information in the embedded spaces. In contrast, this paper explicitly establishes the CFs for both generated and real joint distributions. By inspecting that the CF always exists and uniquely corresponds to one distribution, we propose to calculate the difference between CFs as a vehicle to indicate the discrepancy of joint distributions. However, the calculation of CFs requires excessively sampling in the complex domain, which is prohibitive to learn distributions of images that reside in high dimensions. We thus develop the neural network as a proxy to calculate an upper bound of the CF difference, called neural CF (NCF) metric. Based on the NCF, we establish the CCF-GAN by explicitly modelling the conditional distribution from the joint distribution, allowing for a classified treatment on the image and auxiliary information at different semantic levels. Consequently, the superior performances of our CCF-GAN are verified on both synthetic and real-world datasets.

2. Related Work

cGANs basically optimise joint distributions between images and auxiliary information, which fundamentally differ from unconditional GANs that solely optimise image distributions. The joint optimisation of cGANs allows for controllable generation, the key technique in many scenarios including categorical generation and style transfer. Incorporating the auxiliary information within joint distributions has also been proved to further improve the generation quality, of which the existing cGANs are reviewed in the following.

cGANs by conditioning on $p(x|y)$: The first cGAN [31] proposed to learn the joint distribution by $p(x, y) = p(y)p(x|y)$, and concatenated the auxiliary information $Y$ with the data $X$ as the input of the generator and discriminator, such that the generation and discrimination processes are both informed by the auxiliary information. Similarly, Laplacian pyramid (LAP) GAN [10] and temporal GAN [44] also concatenated $Y$ to $X$ as the input of discriminator, to address the conditional distribution $p(x|y)$. However, since the data $X$ and auxiliary information $Y$ are at different semantic levels, directly concatenating them together may encounter a mismatched information aggregation, leading to instability and inefficiency in training [28, 33]. To relieve this issue, follow-up works [42, 51, 39] proposed to embed $Y$ to certain hidden layers of the discriminator, such that high-level cues of the data have been extracted and then aggregated by the embedded $Y$. Unfortunately, the above methods are designed for applying GAN to accomplish specific tasks such as text-to-image translation [42, 51] and image editing [39].

cGANs by conditioning on $p(y|x)$: Another main trend of cGANs is to decompose $p(x, y)$ into $p(y|x)p(x)$, whereby $p(y|x)$ is predicted by either an implicit or explicit classifier. As one of the representative classifier-free methods, the projection-cGAN was proposed to calculate the likelihood ratios and to indicate $p(y|x)$ by projections, such that the optimisation was implemented under the cross-entropy loss with the theoretical completeness [33]. Due to its simplicity and theoretical beauty, the projection-cGAN has been widely applied in many advanced models, including spectrum normalisation GAN [32], BigGAN [4] and self-attention GAN [50], whereby recent advances including cooperate initialisation [49, 53], knowledge distillation [8]
and gradient regularisation [12]. On the other hand, it has been verified that adding a classifier may improve the generation performance [5]. Auxiliary classifier GAN (ACGAN) is one of the most widely employed cGANs with an explicit classifier, which is trained by the marginal distribution and prediction accuracy [37]. However, ACGAN has been criticised by its behaviour of learning biased distributions, which leads to mode collapse especially when training with large amount of auxiliary information [47, 15, 17]. Later improvements therefore include using a twin auxiliary classifier (TAC) in TACGAN [15], training with a contrastive loss in ContraGAN [20], adding an auxiliary discriminative classifier (ADC) in ADCGAN [19] and implementing several regularisations for stable training in ReACGAN [21]. However, all the above cGANs are based on the cross-entropy loss, which suffer from the incomplete comparisons between two well-separated distributions [2] and may result into mode collapse and instability in training.

**IPM-cGANs:** The IPM has been widely employed for unconditional generation, which successfully reformulates the cross-entropy loss (of predicting real and generated samples) into a theoretically complete distance metric. Notable IPM-GANs include Wasserstein GAN [3], Fisher GAN [34], maximum mean discrepancy GAN [30] and CF-related GANs [1, 30]. To the best of our knowledge, although being explored in some GANs such as the population maximum mean discrepancy GAN [30] and CF-related GANs, applying IPM to cGANs is still yet to start. This is due to the fact that their IPMs are established based on unconditional generation, and the extension to conditional generation has to concatenate the data and auxiliary information together, such that the unconditional settings can be applied. This, however, significantly limits the power of cGANs because it has been verified that decomposing the joint distribution into marginal and conditional distributions can witness remarkable improvements [28, 33]. We also noticed several cGANs tried to combine cross-entropy prediction and IPMs in an *ad hoc* manner [34, 44, 33], which still suffer from the unstable training.

### 3. Methodology

#### 3.1. CF Discrepancy

The CF uniquely defines a random variable $\mathcal{V} \in \mathbb{R}^d$ in terms of cumulative density function (cdf) $F_\mathcal{V}(\mathcal{v})$, given by

$$
\Phi_\mathcal{V}(\mathbf{t}) = \mathbb{E}_\mathcal{V}[e^{j\mathbf{t}^T \mathbf{v}}] = \int_\mathcal{V} e^{j\mathbf{t}^T \mathbf{v}} dF_\mathcal{V}(\mathbf{v}),
$$

where $\mathbb{E}_\mathcal{V}[\cdot]$ denotes the expectation of $\mathcal{V}$. The CF always exists for arbitrary random variables, even when the pdf is not well-defined (for example, the Cantor distribution). When the pdf of a random variable exists, the CF can be formulated as an inverse Fourier transform of $p_\mathcal{V}(\mathbf{v})$, i.e.,

$$
\Phi_\mathcal{V}(\mathbf{t}) = \int_\mathcal{V} e^{j\mathbf{t}^T \mathbf{v}} p_\mathcal{V}(\mathbf{v}) d\mathbf{v}.
$$

In problems including density estimation and generative modelling, the distribution of random variable $\mathcal{V}$ is typically unknown whilst only a set of independent and identically distributed (i.i.d) samples $\{\mathbf{v}_i\}_{i=1}^n$ from $\mathcal{V}$ is available; this prohibits continuous integral over $F_\mathcal{V}(\mathbf{v})$ in CF calculation. Alternatively, we resort to the empirical CF (ECF) that can be calculated as $\Phi_\mathcal{V}(\mathbf{t}) = \frac{1}{n} \sum_{i=1}^n e^{j\mathbf{t}^T \mathbf{v}_i}$, which is an unbiased and consistent estimator of the population $\Phi_\mathcal{V}(\mathbf{t})$ in (1) [13], thus promising a well-defined proxy to approximate the unknown distribution $\mathcal{V}$.

Another appealing property of CF is its boundness, where

$$
|\Phi_\mathcal{V}(\mathbf{t})| = |\int_\mathcal{V} e^{j\mathbf{t}^T \mathbf{v}} dF_\mathcal{V}(\mathbf{v})| \leq \int_\mathcal{V} |e^{j\mathbf{t}^T \mathbf{v}}| dF_\mathcal{V}(\mathbf{v}) = 1,
$$

and reaches its maxima at $\Phi_\mathcal{V}(0) = 1$. In other words, two distributions, $\mathcal{V}$ and $\tilde{\mathcal{V}}$, are automatically aligned in their CFs. It is the fact that comparing two distributions by their pdfs may suffer from misalignment in optimisations, where vanishing gradients and unstable training may exist [2]. This issue motivates the usage of Wasserstein distance, at the cost of increased computational complexity [25] or additional issue motivates the usage of Wasserstein distance, at the cost of increased computational complexity [25] or additional constraints [3]. In contrast, comparing two CFs is naturally resistant to the misalignment issue, whilst enjoying computational ease. We thus use the following $l_2$-norm discrepancy measurement to compare two distributions (i.e., $\mathcal{V}$ and $\mathcal{\tilde{V}}$) via their CFs, on the basis of the uniqueness between a random variable and its CF,

$$
D^2_T(\mathcal{V}||\mathcal{\tilde{V}}) = \int (\Phi_\mathcal{V}(\mathbf{t}) - \Phi_{\mathcal{\tilde{V}}}(\mathbf{t}))(\Phi^*_\mathcal{V}(\mathbf{t}) - \Phi^*_{\mathcal{\tilde{V}}}(\mathbf{t})) p_T(\mathbf{t}) d\mathbf{t}.
$$

In (3), $\Phi^*$ denotes the complex conjugate of $\Phi$, and $p_T(\mathbf{t})$ represents the distribution of $\mathbf{t} \sim T$ that is able to indicate the discrepancy between $\Phi_\mathcal{V}(\mathbf{t})$ and $\Phi_{\mathcal{\tilde{V}}}(\mathbf{t})$ [30]. It has been proved that when the support of $p_T(\mathbf{t})$ resides in $\mathbb{R}^d$, $D$ is a valid distance metric to compare two distributions [30]. We may need to point out that besides the $l_2$ norm, the discrepancy measurement $d(\Phi_\mathcal{V}(\mathbf{t}), \Phi_{\mathcal{\tilde{V}}}(\mathbf{t}))$ can be flexibly chosen by other forms, such as $l_1$ norm or log operation.

Furthermore, we focus on the scenario where $\mathcal{V}$ and $\mathcal{\tilde{V}}$ can be only accessed by their discrete random samples, e.g., $\{\mathbf{v}_i\}_{i=1}^n \sim \mathcal{V}$ for real images and $\{\mathbf{\tilde{v}}_i\}_{i=1}^n \sim \mathcal{\tilde{V}}$ for generated images in image generation tasks. Thus, their CFs can be only accessed by the ECFs, which basically falls into the scope of two-sample test problem, and under mild conditions, the equivalence between two ECFs almost surely (a.s.) ensures the equivalence of two distributions with statistical significance [14], thus indicating the consistency between the corresponding two CFs. Due to this equivalence, instead of using the extra notation $\hat{\Phi}_\mathcal{V}(\mathbf{t})$, we denote in the sequel the ECF by $\Phi_\mathcal{V}(\mathbf{t})$ for simplicity without ambiguity.

More importantly, for conditional generation that involves two joint distributions, for example, $(\mathcal{X}, \mathcal{Y})$ for real images...
and labels, together with \((\tilde{X}, \tilde{Y})\) for generated ones, we are able to formulate \(V = (X, Y)\) and \(\tilde{V} = (X, \tilde{Y})\). This way, the above desirable properties including universal existence and uniqueness still hold for their corresponding ECFS, because \(\Phi_{X, Y}(t) = \Phi_{V}(t)\) and \(\Phi_{X, \tilde{Y}}(t) = \Phi_{\tilde{V}}(t)\). Then, by sampling \(\{t_i\}_{i=1}^{k}\) from \(T\) in (3), we are able to calculate the difference between the two distributions in practice:

\[
D^2_{\Psi}(V||\tilde{V}) = \frac{1}{k} \sum_{i=1}^{k} \left( \Phi_{V}(t_i) - \Phi_{\tilde{V}}(t_i) \right) \left( \Phi^e_{V}(t_i) - \Phi^e_{\tilde{V}}(t_i) \right)
\]

\[
= \frac{1}{k} \sum_{i=1}^{k} \left( \Phi_{X, Y}(t_i) - \Phi_{X, \tilde{Y}}(t_i) \right) \left( \Phi^e_{X, Y}(t_i) - \Phi^e_{X, \tilde{Y}}(t_i) \right)
\]

\[
= D^2_{\Psi}(X||\tilde{X}, \tilde{Y}), \tag{4}
\]

where \(\Phi_{X, Y}(t_i)\) and \(\Phi_{X, \tilde{Y}}(t_i)\) represent the ECFs for real and generated joint distributions, respectively. It should be pointed out that in (4), the number of samples \(k\) plays a crucial role in distinguishing \((X, Y)\) from \((X, \tilde{Y})\), so as to indicate sufficient discrepancy for probability estimation. We illustrate in Fig. 1 that without any discriminator modules, optimising a generator network solely by setting \(k = 128\) and \(T\) to be the standard Gaussian distribution in (4) can generate roughly realistic images towards MNIST digits [27].

However, the grey-scale digital images from MNIST dataset [27] with size \(28 \times 28\) are simplified scenarios when comparing with real-world images. When optimising images of high dimensions and with diversifying content, \(k\) has to increase exponentially, especially for high-dimensional data, encountering the curse of dimensionality (cod) problem. To address this, rather than Gaussian distribution, \(\{t_i\}_{i=1}^{k}\) need to be smartly chosen. More importantly, in this preliminary experiment of Fig. 1, we straightforwardly concatenated the label information \(Y\) with the images \(X\), which has been verified to be ineffective since the pixel-wise images and class-wise labels are essentially at different semantic levels [28, 33]. In Section 3.2, we first introduce the way of addressing the cod problem, followed by a novel way to treat the semantic levels at different importance in Section 3.3.

### 3.2. Adversarial NCF Learning

To address the cod problem when calculating the discrepancy between ECFS, several methods [1, 30], which were born for unconditional generation, proposed to reduce the dimensions of images by learning an embedding function \(f(\cdot) : \mathbb{R}^d \rightarrow \mathbb{R}^{d'}\); where \(d' \leq d\) [1, 30]; this allows for explicitly enumerating \(T\) in the low dimension \(d'\) when comparing two embedded distributions \(f(V) \in \mathbb{R}^{d'}\) and \(f(\tilde{V}) \in \mathbb{R}^{d'}\). However, the embedding requires extra requirements on the function \(f(\cdot)\), including injection [1] and bijection [30], resulting into additional hyper-parameters and instability when training GANs. More importantly, the embedding function \(f(\cdot)\) is basically implemented by the discriminator network (also known as the critic), which is highly non-linear. Its extension to conditional generation is therefore highly limited, and the only possible compromise is to embed a concatenated joint distribution \(f(V) = f(X, Y)\), the way that the majority of IPM-cGANs operate. This compromise, however, is proved to be inefficient [28, 33].

On the other hand, the very basic operation of CFs in (1), namely, \(t^T v\), projects from high dimension \(\mathbb{R}^d\) to a scalar \(\mathbb{R}\). Thus, instead of explicitly enumerating \(T\), we propose to implicitly optimise \(T\) when comparing two complicated distributions in (4). This is also motivated by the Cramer-Wold Theorem [7], which states that two random variables \(V, \tilde{V} \in \mathbb{R}^d\) have the same distribution if and only if distributions of \(t^T V \in \mathbb{R}\) and \(t^T \tilde{V} \in \mathbb{R}\) are the same for all \(t \in \mathbb{R}^d\). In other words, we are able to compare two complicated and high-dimensional distributions, by means of their infinite projections \(\mathbb{R}^d \rightarrow \mathbb{R}\) in the one-dimensional space. Therefore, instead of explicitly sampling \(T\) from several pre-defined distributions [1, 30], we propose to implicitly search all possible \(t\) and directly output the corresponding projections \(t^T V\) and \(t^T \tilde{V}\), so as to compare the projected distributions in low dimensions. Thus, the difference of two CFs in (4) can be generalised by the projection function \(f\) as

\[
D_{\mathcal{F}}(V||\tilde{V}) = \left( \frac{1}{k} \sum_{i=1}^{k} \left( \Phi^i_V - \Phi^i_{\tilde{V}} \right) \left( \Phi^i_{V} - \Phi^i_{\tilde{V}} \right) \right)^{\frac{1}{2}}, \tag{5}
\]

where \(\Phi^i_V\) is calculated by the \(i\)-th projection \(f_i(\cdot)\):

\[
\Phi^i_V = \mathbb{E}_V [e^{f_i(V)}] = \frac{1}{n} \sum_{v_i} e^{f_i(v_i)}. \tag{6}
\]

In (5), \(f_i(v_i)\) is parameterised by the proposed NCF network, whereby the input is \(v_i\) and \(f_i(v_i)\) represents the \(i\)-th dimension output of the NCF network.

Furthermore, compared with excessively sampling by varying \(f(v)\), it is more efficient to decide the "best repre-
sentative\(^*\) samples that are able to maximally distinguish the two CFS in \( D_{\mathcal{X}}(\mathcal{V}||\mathcal{V}) \) in (5), as follows

\[
\mathcal{L}(\mathcal{V}||\mathcal{V}) = \max_f D_{\mathcal{X}}(\mathcal{V}||\mathcal{V}). \tag{7}
\]

**Lemma 1.** For any two random variables \( \mathcal{V}, \mathcal{W} \in \mathbb{R}^d \), \( \mathcal{L}(\mathcal{V}||\mathcal{W}) \geq D_T(\mathcal{V}||\mathcal{W}) \) for any \( T \), where \( D_T(\mathcal{V}||\mathcal{W}) \) is defined in (4).

Lemma 1\(^1\) proves an upper bound of \( \mathcal{L}(\mathcal{V}||\mathcal{W}) \) against the true CF discrepancy \( D_T(\mathcal{V}||\mathcal{W}) \). This way, minimising \( \mathcal{L}(\mathcal{V}||\mathcal{W}) \) naturally reduces the difference between two distributions, as measured by \( D_T(\mathcal{V}||\mathcal{W}) \). We further provide in Lemma 2 that the measurement \( \mathcal{L}(\mathcal{V}||\mathcal{W}) \) is a valid distance metric, which is able to precisely reflect the difference between two distributions.

**Lemma 2.** If \( \mathcal{V}, \mathcal{W} \in \mathbb{R}^d \) are two random variables, \( \mathcal{L}(\mathcal{V}||\mathcal{W}) \) in (7) is a valid distance metric.

### 3.3. Conditional Generation by CCF-GAN

By far, the conditional generation can be achieved by setting \( \mathcal{V} = (\mathcal{X}, \mathcal{Y}) \) as illustrated by Fig. 1. However, since the image \( \mathcal{X} \) and auxiliary information \( \mathcal{Y} \) reside at different semantic levels, directly stacking them together is problematic in cGANs [28, 33]. In this section, we propose to treat them separately such that the auxiliary information \( \mathcal{Y} \) can be well accommodated along with generating \( \mathcal{X} \), thus enjoying improved generation performances. More specifically, the definition of CF allows for an explicit decomposition on \( p(x) \) and \( p(y|x) \) as follows,

\[
\Phi_{\mathcal{V}}(t) = \Phi_{\mathcal{X},\mathcal{Y}}(t) = \int_x \int_y e^{j(t^T_x x + t^T_y y)} p(x, y) dx dy
\]

\[
= \int_x \left[ \int_y e^{jT^T_y y} p(y|x)dy \right] e^{jT^T_x x} p(x) dx,
\]

where \( t = [t^T_x, t^T_y]^T \). We may need to point out that (8) plays a key role in our CCF-GAN, which effectively decomposes \( \mathcal{Y} \) from \( \mathcal{X} \). In many tasks, the auxiliary information follows the discretely distribution, e.g., the class labels. Thus, we are able to obtain the CF of \( p(y|x) \) in (8) as

\[
\int_y e^{jT^T_y y} p(y|x)dy = \sum_{y=1}^{c} e^{jT^T_y y_i} p(y_i|x),
\]

where \( c \) is the number of discrete values of \( \mathcal{Y} \). Correspondingly, the ECF of \( (\mathcal{X}, \mathcal{Y}) \) now arrives at

\[
\Phi_{\mathcal{X},\mathcal{Y}}(t) = \frac{1}{n} \sum_{i=1}^{n} \sum_{y=1}^{c} e^{jT^T_y y_i} p(y_i|x_i) e^{jT^T_x x_i}. \tag{10}
\]

More importantly, the image distribution \( \mathcal{X} \) typically resides in high dimensions and thus requires smart strategies to avoid

---

\(^1\)Please refer to the supplementary material for the proofs of all lemmas.
The proposed CCF-GAN is analysed by illustrating the dynamics of both $e^{i\lambda^T x}$ and $e^{i\lambda^T y}$ for conditional generation.

In contrast, our CCF-GAN, benefiting from proposing the NCF network to directly output $\hat{x}$, is able to explicitly extract $\lambda$ from the joint distribution, and the remaining part is formulated by the conditional distribution $p(y|x)$. This way, the data distribution and auxiliary information can be well learned with different importance, allowing for an optimised discrepancy measure between the real $(\lambda, y)$ and generated $(\hat{\lambda}, \hat{y})$ joint distributions. In practice, we implement the proposed NCF network as the discriminator, so as to measure the generated distribution $(\hat{\lambda}, \hat{y})$ from the generator $g(\cdot)$. The details are presented in Algorithm 1, whereas the pipeline is provided in Fig. 2. We further illustrate the superiority of the proposed CCF-GAN against existing CF related GANs in Fig. 2. As can be seen from this figure, CCF-GAN can well separate real and generated samples at the middle stage, whereby RCFGAN fails. At the end stage, real and generated samples are aligned by CCF-GAN, whereas separation still exists in OCFGAN.

4. Experiment

4.1. Experimental Settings

Datasets: By comparing the proposed CCF-GAN with other state-of-the-art cGANs, we performed the experiments to evaluate the performances of conditional generation on 1 synthetic dataset and 3 widely accepted real-world datasets, namely, CIFAR10 [26], VGGFace2 [38] and ImageNet [43]. For the synthetic dataset, we employed a mixture of 3 von Mises–Fisher (vMF) distributions [6], and their parameters
\{p, \tau, \theta\} were set to \{0.33, 30, 2\pi/3\}, \{0.33, 30, 4\pi/3\} and
\{0.33, 30, 2\pi\}, respectively, where 100k points were ran-
domly sampled. The real samples were plotted in Fig. 3-(e).
More importantly, the consideration of using vMF clusters is because the vMF distribution is basically supported in low dimensions, which can effectively mimic the real-world sce-
narios where the data are typically in high dimensions and
the generating spaces reside on low dimensions. For real-
world scenarios, images in CIFAR10 dataset were of size
32 \times 32. We followed [15] to randomly select 200, 500 and
1,000 classes from the VGGFace2 dataset, denoted as VGG-
Face_c200, VGGFace_c500, and VGGFace_c1000. Then,
the images were centercropped and resized to 64 \times 64. For
ImageNet, we resized the images to resolution of 128 \times 128.

**Metrics:** The widely applied Fréchet inception distance
(FID) [18] metric was adopted in our evaluation to assess the
generation quality of GANs, which basically implements the
Wasserstein distance between the real and generated features
witnessed further improvements on the StudioGAN plat-
form. We sampled world scenarios, images in CIFAR10 dataset were of size
(FID) [18] metric was adopted in our evaluation to assess the
stability was also evaluated by repeatedly training GANs
under various conditions.

**Baselines:** We compared the proposed CCF-GAN with the
BigGAN [4], ACGAN [37], TACGAN [15] and ADCGAN
[19]. Besides, FisherGAN [34] and cRCFGAN [29] were
adopted for comparison as conditional IPM-GANs. Fur-
thermore, we also evaluated our CCF-GAN with several
most recent GANs without the classifier, including Contra-
GAN [20], KD-DLGAN [8] and DigGAN [12]. We
implemented our CCF-GAN on the Pytorch BigGAN platform3,
by using exactly the same architecture for the generator and
discriminator networks as the BigGAN [4]. All the compar-
ing cGANs were trained and tested based on the Pytorch
BigGAN platform, under the same architecture. Most
recently, there comes with a new rising-star platform called
the StudioGAN4 [22], which implements ContraGAN [20]
and ReACGAN [21]. Because the new StudioGAN em-
ployed random flipping and different image resize functions
by default, we believe it is unfair to report the result upon
the StudioGAN platform. Otherwise, it might be unclear
to show the origin of our improvements. Indeed, we have
witnessed further improvements on the StudioGAN plat-
form on all the datasets, which we decided to put in the

---

Table 1: Comparison on FID scores against existing state-of-the-art baselines in Tables 1 and 2. As can be seen from Table 1, the proposed CCF-GAN achieved the lowest (best)

<table>
<thead>
<tr>
<th>Method</th>
<th>CIFAR10 VGGFace_c200 VGGFace_c500 VGGFace_c1000</th>
</tr>
</thead>
<tbody>
<tr>
<td>BigGAN [4]</td>
<td>14.73</td>
</tr>
<tr>
<td>ACGAN [37]</td>
<td>8.01</td>
</tr>
<tr>
<td>FisherGAN [34]</td>
<td>11.46</td>
</tr>
<tr>
<td>TACGAN [15]</td>
<td>8.42</td>
</tr>
<tr>
<td>cRCFGAN [29]</td>
<td>6.90</td>
</tr>
<tr>
<td>ContraGAN [20]</td>
<td>10.60†</td>
</tr>
<tr>
<td>ADCGAN [19]</td>
<td>7.17</td>
</tr>
<tr>
<td>DigGAN [12]</td>
<td>8.49†</td>
</tr>
<tr>
<td>CCF-GAN (Ours)</td>
<td>6.08</td>
</tr>
</tbody>
</table>

† denotes that the results are reported from the corresponding paper, whereas ‡ from [15]. Otherwise, we ran the available codes by the corresponding default settings. We denote the best FID by red color and the second best by blue color.

Technical details: In our experiments, we selected a steady learning rate of 0.0001 for generator and 0.0002 for discriminator with classifier. Although being able to achieve conditional generation by directly inputting ground-truth labels, the default setting of our CCF-GAN included the classifier [15]. The discriminator, together with the classifier, was trained 2 steps per generator update. For other comparing methods that were replicated by their public repositories, we set the same hyper-parameters as those in the corresponding papers. More importantly, the batch sizes for CIFAR10, VGGFace_c200, VGGFace_c500 and VGGFace_c1000 were set to 64. Batch size for ImageNet was set to 256.

4.2. Distribution Fitting Results on Synthetic Data

We illustrate in Fig. 3 the comparisons among the AC-
GAN, TACGAN, ADCGAN and the proposed CCF-GAN,
on the 2D synthetic dataset. As can be seen from this fig-
ure, our CCF-GAN almost recovered the ground-truth dis-
tribution, whereas others are either over-concentrated (AD-
CGAN) or imbalanced (ACGAN and TACGAN). This vali-
dates the effectiveness of employing the NCF in our CCF-
GAN, which stably and accurately measured two distribu-
tions even when they were supported in low dimensions.
In contrast, existing cGANs that are designed based on the
cross-entropy loss may suffer from the ill-posed discrepancy
measure, such that the fitted distributions were biased.

4.3. Realistic Image Generation Results

We also compared our CCF-GAN with existing state-of-
the-art baselines in Table 1 and 2. As can be seen from
Table 1, the proposed CCF-GAN achieved the lowest (best)

---

2 Due to non-linear coupling, RCFGAN was designed for unconditional generation. cRCFGAN is its compromised variant by channel-wise concatenating images and labels as augmented input for condition generation.
3 https://github.com/ajbrock/BigGAN-PyTorch
4 https://github.com/POSTECH-CVLab/PyTorch-StudioGAN
Figure 3: Distribution fitting results on 2D synthetic dataset, which consists of 100k samples from the mixture of \(\nu\)MF distributions. Please note that ACGAN, TACGAN, ADCGAN and CCF-GAN were trained by the same networks, which consist of 4-layer (for generator) and 3-layer (for discriminator) fully connected neural networks of hidden size equal to 10.

Figure 4: Conditional image generation on CIFAR10, VGGFace2_c1000 and ImageNet datasets by the proposed CCF-GAN. Each row represents one class-conditioned generation.

Table 2: Comparison on the ImageNet. Symbol * denotes that the results are reported from [19], whereas † from [15], ‡ from [21] and ** for [20]. Otherwise, we ran the available codes by the corresponding default settings.

<table>
<thead>
<tr>
<th>Method</th>
<th>ImageNet</th>
<th>FID</th>
<th>IS</th>
</tr>
</thead>
<tbody>
<tr>
<td>BigGAN [4]</td>
<td></td>
<td>22.77</td>
<td>38.05†</td>
</tr>
<tr>
<td>ContraGAN [20]</td>
<td></td>
<td>19.69**</td>
<td>31.10†</td>
</tr>
<tr>
<td>ACGAN [37]</td>
<td></td>
<td>184.41†</td>
<td>7.26†</td>
</tr>
<tr>
<td>TACGAN [15]</td>
<td></td>
<td>23.75†</td>
<td>28.86†</td>
</tr>
<tr>
<td>ReACGAN [21]</td>
<td></td>
<td>13.98†</td>
<td>68.27‡</td>
</tr>
<tr>
<td>ADCGAN [19]</td>
<td></td>
<td>16.75*</td>
<td>55.43*</td>
</tr>
<tr>
<td>CCF-GAN (Ours)</td>
<td></td>
<td><strong>11.34</strong></td>
<td><strong>180.84</strong></td>
</tr>
</tbody>
</table>

FID against all the compared methods. Similar results can be also concluded in Table 2, whereby the proposed CCF-GAN achieved the value 11.34 of FID by training under the batch size of 256 for the ImageNet dataset. The IS score of our CCF-GAN, however, was much remarkable and reached 180.84, almost tripled against the second best ReACGAN.

We further present in Fig. 4 the conditional generation results of our CCF-GAN. As can be seen from this figure, our CCF-GAN achieved high-quality image generation. More importantly, by inspecting each row, the class-wise semantics are obvious and the generated images within each class are of diversifying content, which verifies that the proposed CCF-GAN, by incorporating the CF distance measure, is able to overcome the mode collapse issue. In Fig. 5, the interpolation was performed across different classes, whereby the interpolation between two different faces is smooth, verifying the desirable continuity of the latent space learnt by our CCF-GAN. Qualitative comparisons, together with more subjective results and analysis, are provided in the supplementary material.

Figure 5: Interpolation across class labels of the proposed CCF-GAN, which was trained on VGGFace_c1000 dataset.

4.4. In-depth Analysis

Ablation study on \(k\), classifier and \(t_y\): Since the number of \(t\) samples, namely, \(k\), plays a crucial role in distinguishing CFs between generated and real distributions, FIDs of varying \(k\) are plotted in Fig. 6-(a). We thus can conclude that in complicated real-world scenarios, the proposed NCF effectively resolves cod issue, and \(k = 256\) is sufficient to be the best among the existing baselines. Another ablation
investigates the usage of classifier, as shown in Fig. 6-(b).
As can be seen from this figure, our CCF-GAN can still achieve conditional generation by directly using the ground-truth labels as \(p(y|x_i)\) in (11), i.e., without the classifier. However, training a classifier witnessed improvements on FIDs in our CCF-GAN, which is in accordance with [5]. Moreover, we also ablated on different choices on \(t_y\), including fixed linear space rule of range \([-\alpha, \alpha]\), as well as random samples from uniform and Gaussian distributions. Table 3 indicates that the proposed CCF-GAN performs well under \(t_y\) from different distributions, particularly when \(t_y\) was fixed to be \([-1, 1]\).

Analysis on mode collapse: We quantitatively evaluated the mode collapse of generation by the precision and recall metrics [45] plotted in Fig. 7. From this figure, we can find that our CCF-GAN achieved the highest (best) precision and recall values, and the improvements on the recall were even more significant, which verifies the capability of relieving mode collapse of our CCF-GAN.

![Figure 6: Ablation study on CIFAR dataset. (a) Training CCF-GAN by varying \(k\). (b) Training CCF-GAN with and without the classifier.](image)

![Figure 7: Precision and recall metrics [45] on CIFAR10 and VGGFace_c1000 datasets.](image)

![Figure 8: Stability evaluation under 48 conditions. The horizontal axis represents FID thresholds, and vertical axis denotes the number of trials whose best FIDs are lower than the corresponding FID threshold within 100\(k\) iterations.](image)

Table 3: Ablations on different choices on \(t_y\).

<table>
<thead>
<tr>
<th>(\alpha)</th>
<th>Fixed (t_y)</th>
<th>Uniform (t_y)</th>
<th>Gaussian (t_y)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\text{FID} \downarrow)</td>
<td>6.51</td>
<td><strong>6.08</strong></td>
<td>7.02</td>
</tr>
<tr>
<td></td>
<td>7.05</td>
<td>7.07</td>
<td></td>
</tr>
</tbody>
</table>

5. Conclusion

In this paper, we have proposed a novel CCF-GAN for consistently improving the conditional generation performances on both synthetic and real-world datasets. Different from the existing cGANs built upon the cross-entropy loss, our CCF-GAN benefits from the characteristic function (CF), which processes unique and universal correspondence to a random variable, even when the random variable does not possess probability density function. On the basis of the CF, we have proposed an efficient neural characteristic function (NCF) network to calculate the difference between CFs with theoretical completeness. We further explicitly decomposed the joint distribution by the marginal and conditional distributions, with classified treatment for different semantics levels. This way, CCF-GAN has overcome the deficiency of almost all cGANs of employing the cross-entropy loss. The experimental results have verified that the proposed CCF-GAN achieved the best conditional generation, whilst significantly reducing mode collapse and instability in cGANs.

Acknowledgments. This work was supported by NSFC under Grants 62206011, 62250001 and 62231002, and Beijing Natural Science Foundation under Grant JQ20020 and L223021.
References


