applied sciences

MDPI

Article

Express Construction for GANs from Latent Representation to Data Distribution

Minghui Liu †, Jiali Deng †, Meiyi Yang, Xuan Cheng, Tianshu Xie, Pan Deng, Xiaomin Wang and Ming Liu *

School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China; minghuiliu@std.uestc.edu.cn (M.L.); dengjiali@std.uestc.edu.cn (J.D.); meiyiyang@std.uestc.edu.cn (M.Y.); cs_xuancheng@std.uestc.edu.cn (X.C.); xietianshucs@163.com (T.X.); 201922080828@std.uestc.edu.cn (P.D.); xmwang@uestc.edu.cn (X.W.)

* Correspondence: csmliu@uestc.edu.cn

† These authors contributed equally to this work.

Featured Application: This concise is a novel training methodology for GANs with strong generalization ability, speed the training up, and less prone to mode collapse.

Abstract: Generative Adversarial Networks (GANs) are powerful generative models for numerous tasks and datasets. However, most of the existing models suffer from mode collapse. The most recent research indicates that the reason for it is that the optimal transportation map from random noise to the data distribution is discontinuous, but deep neural networks (DNNs) can only approximate continuous ones. Instead, the latent representation is a better raw material used to construct a transportation map point to the data distribution than random noise. Because it is a low-dimensional mapping related to the data distribution, the construction procedure seems more like expansion rather than starting all over. Besides, we can also search for more transportation maps in this way with smoother transformation. Thus, we have proposed a new training methodology for GANs in this paper to search for more transportation maps and speed the training up, named Express Construction. The key idea is to train GANs with two independent phases for successively yielding latent representation and data distribution. To this end, an Auto-Encoder is trained to map the real data into the latent space, and two couples of generators and discriminators are used to produce them. To the best of our knowledge, we are the first to decompose the training procedure of GAN models into two more uncomplicated phases, thus tackling the mode collapse problem without much more computational cost. We also provide theoretical steps toward understanding the training dynamics of this procedure and prove assumptions. No extra hyper-parameters have been used in the proposed method, which indicates that Express Construction can be used to train any GAN models. Extensive experiments are conducted to verify the performance of realistic image generation and the resistance to mode collapse. The results show that the proposed method is lightweight, effective, and less prone to mode collapse.

Keywords: generative adversarial networks; mode collapse; transportation maps; latent representation; express construction

Citation: Liu, M.; Deng, J.; Yang, M.; Cheng, X.; Xie, T.; Deng, P.; Wang, X.; Liu, M. Express Construction for GANs from Latent Representation to Data Distribution. *Appl. Sci.* **2022**, *12*, 3910. https://doi.org/10.3390/app12083910

Academic Editors: Andrea Prati, Luis Javier García Villalba and Vincent A. Cicirello

Received: 14 March 2022
Accepted: 10 April 2022
Published: 13 April 2022

1. Introduction

Generative methods are one of the most promising approaches toward automatically learning features from a given high-dimension data distribution and then producing new samples approximating the truth [1]. Currently, the most prominent approaches are auto-regressive models [2,3], variational auto-encoders (VAEs) [4], generative adversarial networks (GANs), and a unifying framework combining the best of VAEs and GANs like Wasserstein auto-encoders (WAE) [5]. They all have significant strengths and weakness, and among them, GANs have the most significant impact. GANs always consist of two networks: a generator network, which maps a random noise ζ to a data distribution P_r of

real ones without calculating the sample likelihood, and a discriminator network trained to assess whether the input sample from the data distribution. However, this formulation has multiple potential problems, such as training instability and mode collapse, especially the latter, which is the most urgent and difficult problem in GANs.

Generators and Discriminators in GANs are always composed of deep neural networks, which can only represent continuous mappings. However, as pointed out by these works, the transport maps may be discontinuous if there are too many modes in the data distribution. This intrinsic conflict will lead to mode collapse or mode mixture [6–8]. Although a number of alternatives that can increase diversity have been proposed, the problem remains unsolved.

In deep learning, the manifold distribution hypothesis is well accepted, which assumes the distribution of a specific class of natural data is concentrated on a low dimensional manifold embedded in the high dimensional data space [9]. Auto-encoders (AEs) have demonstrated the capability of learning a subspace for dimensionality reduction, which consistently attempts to find the encoding maps between the data manifold embedded in the image space, and the decoding maps between the data manifold embedded in the latent space [10]. Generally, the data distribution will be embedded into the latent space Ω in AEs. Compared to random noises, the latent representation sampled from the latent space is a low-dimensional representation mapped from the real data distribution. Both the transportation map from that to the data distribution, and the one from random noise to the latent representation, are easier to search and construct. Note that this implies that it could ease up the intrinsic conflict. Thus, we tackle the mode collapse problem for GANs by decomposing the generating procedure into two phases: the generation from random noise to the latent representation and the generation from the latent representation to the data distribution.

In this paper, we have proposed a training methodology for GANs named Express Construction. An auto-encoder is trained with mean absolute error (MAE) loss first to embed the image manifold $\mathcal{X}$ into the latent manifold Ω. The distribution sampled from Ω as the latent representation serves as the target of the first stage. To this end, a standard GAN model with a small scale is trained to map random noises into the latent representation by creating new samples that are intended to come from the same distribution as the training data. Then, the output of this model will be used as the input for the next stage. Another large-scale GAN model is trained to generate new samples close to the data distribution from the latent representation rather than random noises. We can reduce the task complexity and search for more transportation maps to benefit from indirect construction and smoother transformation. Finally, we provide theoretical steps toward understanding the training dynamics of these procedures and prove our assumptions. Extensive experiments are conducted to verify the performance and contrast with other works. Results show that Express Construction effectively alleviates the mode collapse problem and speeds the training up.

This work provides these primary contributions:

- We have proposed a novel training methodology for GANs to tackle the mode collapse problem, named Express Construction. The generating procedure in the proposed method will be decomposed into two steps: generating the latent representation using random noises, and generating the final results close to the data distribution from the latent representation rather than random noises. To the best of our knowledge, we are the first to search for more transportation maps in this way.
- Theoretical statements are provided to prove our assumptions under the views on the training dynamics. Besides, the transportation cost will be discussed, which indicates Express Construction is lightweight and effective.
- We conduct extensive experiments and evaluate our contributions using different datasets from small to large scale. The results show Express Construction is less prone to mode collapse and is able to generate realistic samples.

2. Related Works

The training for conventional GANs always starts with a random Gaussian distribution. However, general DNNs can only approximate continuous mappings, while the optimal transportation mapping is discontinuous if the target measure is disconnected or just non-convex [6]. Many researchers have observed this drawback and attempted to solve it.

GMAN [11] is a framework of GANs whose internal samples are created from multiple generators, and one of them is randomly selected as the final output. The procedure is similar to the mechanism of a probabilistic mixture model. It can generate diverse and appealing recognizable objects with different resolutions, and specialize in capturing different types of objects by the generators. MAD-GAN [12] is another work similar to GMAN. It also has one discriminator and multiple generators. However, its discriminator must learn to push different generators towards different identifiable modes rather than using classifiers. Therefore, both of them are costly with strong constraints. On the other hand, unrolled GAN [13] is a method to stabilize GANs and increase diversity by defining the generator objective with respect to an unrolled optimization of the discriminator. Obviously, the mode collapse problem is still far from solved.

In particular, AE-OT is an effective model which explicitly separates the manifold embedding and the optimal transportation. It is carried out using an auto-encoder to map the images onto the latent space. Then, a graphics processing unit (GPU)-based convex optimization is used to find the discontinuous transportation maps. Finally, composing the extended optimal transport map and the decode, they can generate new samples from the white noise [7,14]. Though it effectively tackles the mode collapse problem, the visual quality of the generated samples is unsatisfactory. GANs can generate very convincing images, sharper than ones produced by auto-encoders using pixel-wise losses. Therefore, AE-OT-GAN is proposed to combine the advantages of both models and generate high-quality images without mode collapse. Similarly, it can also embed the low dimensional image manifold into the latent space by an auto-encoder, and the semi-discrete optimal transport (SDOT) map is used to generate new latent codes. However, after the two steps, they can directly sample from the latent distribution by applying a piece-wise linear extension map on the uniform distribution to train the GAN model [15]. It can produce more realistic images, but the SDOT map is still needed to be computed for constructing a continuous latent distribution, which is time-costing and may reduce the generating performance.

The most important and direct way to tackle the mode collapse problem is to construct transportation maps more effectively. However, it is challenging with a high computational cost, and cannot search. Instead, searching for relay nodes is a more reliable way. Although some works have been proposed to search for optimal transportation and thus reduce the discontinuous transportation maps, the computational cost is enormous and potentially reduces the robustness of the model. Therefore, the search for a more stable and simple way to search more transportation maps is a significant challenge.

3. Problem Statement

GANs always want to search for a mapping from random noise to the latent space. Unfortunately, the searching will not converge or will converge to one continuous branch of the target mapping, leading to the mode collapse problem. Instead, the mapping from the latent representation to the data distribution is easier to search, and is analogous to the mapping from random noise to the latent representation. That means it is effective to tackle the mode collapse problem and potentially speed the training up. Thus, we have proposed a training methodology for GANs, named Express Construction, and provide mathematical statements as follows.

We will first recall the regularity analysis for mode collapse [6]. The generator map $g_\theta : (\mathcal{Z}, \zeta) \longrightarrow (\sum, \mu_\theta)$ can be further decomposed into two steps,

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

where T is a transportation map, mapping the noise ζ to μ in the latent space $\mathcal{Z}$, g is the manifold parameterization, mapping local coordinates in the latent space to the manifold Σ. That is to say, g provides a local chart of the data manifold, and T realizes the probability measure transformation. By manifold structure assumption, the local chart representation $g : \mathcal{Z} \longrightarrow \Sigma$ is continuous. However, according to the regularity theory of the optimal transportation map, except in very rare situations, the transportation map T is always discontinuous. This intrinsic conflict leads to the mode collapse problem. Instead, Express Construction is able to search for more transportation maps by decomposing the procedure into two phases, making the manifold transformation in Equation (1) smoother rather than a drastic transformation from a low dimension to a high one. We now compare them in terms of training dynamics and total transportation cost, as follows.

We first recall the theory of the perfect discriminator, following [16].

Definition 1. *Let $\mathcal{M}$ and $\mathcal{P}$ be two boundary free regular submanifolds of $\mathbb{R}^d$. Let $x \in \mathcal{M} \cap \mathcal{P}$ be an intersection point of the two manifolds. $\mathcal{M}$ and $\mathcal{P}$ intersect transversally in x if $T_x\mathcal{M} + T_x\mathcal{P} = T_x\mathbb{R}^d$, where $T_x\mathcal{M}$ means the tangent space of $\mathcal{M}$ around x. Accordingly, $\mathcal{M}$ and $\mathcal{P}$ perfectly align if there is an $x \in \mathcal{M} \cap \mathcal{P}$ such that $\mathcal{M}$ and $\mathcal{P}$ do not intersect transversally in x.*

Lemma 1. *Let $\mathcal{M}$ and $\mathcal{P}$ be two regular submanifolds of $\mathbb{R}^d$ that do not perfectly align and do not have full dimension. Let $\mathcal{L} = \mathcal{M} \cap \mathcal{P}$. If $\mathcal{M}$ and $\mathcal{P}$ do not have a boundary, then $\mathcal{L}$ is also a manifold, and has a strictly lower dimension than both the one of $\mathcal{M}$ and the one of $\mathcal{P}$. If they have a boundary, $\mathcal{L}$ is a union of, at most, four strictly lower dimensional manifolds. In both cases, $\mathcal{L}$ measures 0 in both $\mathcal{M}$ and $\mathcal{P}$.*

Lemma 2. *Let $\mathbb{P}_r$ (data distribution) and $\mathbb{P}_g$ (generated distribution) be two distributions that have support contained in two closed manifolds $\mathcal{M}$ and $\mathcal{P}$ that do not perfectly align and do not have full dimension. We further assume that $\mathbb{P}_r$ and $\mathbb{P}_g$ are continuous in their respective manifolds, meaning that if there is a set A with measure 0 in $\mathcal{M}$, then $\mathbb{P}_r(A) = 0$ (and analogously for $\mathbb{P}_g$). Then, there exists an optimal discriminator $D^* : \mathcal{X} \to [0, 1]$ that has an accuracy of 1 and, for almost any x in $\mathcal{M}$ or $\mathcal{P}$, D^*, is smooth in a neighbourhood for x and $\nabla D^*(x) = 0$.*

As mentioned in Lemmas 1 and 2, we can assume that there is a perfect discriminator D that is smooth and constant almost everywhere in $\mathcal{M}$ and $\mathcal{P}$, while both of their supports are disjointed or lie on low dimensional manifolds.

Definition 2. *Let $\mathcal{M}$ and $\mathcal{P}_1$ be two regular submanifolds of $\mathbb{R}^d$ that do not perfectly align and do not have full dimension, and are analogous for $\mathcal{P}_2$. Let $\mathbb{P}_r$ be defined as the data distribution that has support contained in $\mathcal{M}$. Let $\mathbb{P}_g^1$ be defined as the generated distribution of the transportation map from random noise to the data distribution that has support contained in $\mathcal{P}_1$. Then, let $\mathcal{P}_g^2$ be defined as the generated distribution of the transportation maps from the latent representation to the data distribution, and have support contained in $\mathcal{P}_2$. Let $\mathcal{L}_1 = \mathcal{P}_1 \cap \mathcal{M}$ and $\mathcal{L}_2 = \mathcal{P}_2 \cap \mathcal{M}$ be two manifolds that have strictly lower dimension than both $\mathcal{M}$, and one of $\mathcal{P}_1$ and $\mathcal{P}_2$, respectively.*

Significantly, the latent representation is a low-dimensional mapping related to the data distribution, which implies that the training dynamics from $\mathcal{Z}$ to Σ is always more stable than the one from ζ to Σ, and we can prove it as follows.

Proof. Assume that $\mathbb{P}_r$, $\mathbb{P}_g^1$, and $\mathbb{P}_g^2$ are continuous in their respective manifolds. Let $x \in \mathcal{M} \setminus (\mathcal{L}_2 - \mathcal{L}_1)$, thus, $x \in \mathcal{P}_1$ is an open set that is a ball of radius ε_x such that $B(x, \varepsilon_x) \cap \mathcal{P}_1 = \emptyset$, which indicates

$$\hat{\mathcal{M}} = \bigcup_{x \in \mathcal{M} \setminus (\mathcal{L}_2 - \mathcal{L}_1)} B(x, \varepsilon_x \setminus 3) \tag{2}$$

For $\hat{\mathcal{P}}_1$ and $\hat{\mathcal{P}}_2$ analogously.

By construction, these are both open sets on $\mathbb{R}^d$. Since $\mathcal{M} \setminus (\mathcal{L}_2 - \mathcal{L}_1) \subseteq \hat{\mathcal{M}}$ and $\mathcal{P}_1 \setminus (\mathcal{L}_2 - \mathcal{L}_1) \subseteq \hat{\mathcal{P}}_1$, the support of $\mathbb{P}_r$ and $\mathbb{P}_g^1$ is contained in $\hat{\mathcal{M}}$ and $\hat{\mathcal{P}}_1$, respectively. This indicates $\hat{\mathcal{M}} \cap \hat{\mathcal{P}}_1 = \emptyset$. That is to say, $D^*(x) = 1$ for all $x \in \hat{\mathcal{M}}$ and 0 elsewhere, and $\nabla_x D^*|_{\hat{\mathcal{P}}_1} = 0$. Namely, $D^*(x)$ is a perfect discriminator to make the training dynamics unstable.

Instead, we can say that $\nabla_x D^*|_{\hat{\mathcal{P}}_2} \neq 0$ and $\hat{\mathcal{M}} \cap \hat{\mathcal{P}}_2 \neq \emptyset$. This is caused by the regular submanifold of $\mathcal{Z}$ which is undoubtedly higher than ζ, and leads to the transformation from $\mathcal{Z}$ to Σ, and is more continuous than another one. We say that $\mathcal{P}_2 \subseteq \mathcal{P}_1$. Thus, we can say $\nabla_x D^*|_{\hat{\mathcal{P}}_2} \neq 0$. In other words, $\nabla_x D^*$ is not a perfect discriminator for the transformation from $\mathcal{Z}$ to Σ. □

We then compare the total transportation cost between these two procedures, as follows.

According to the Brenier's theorem [14], the transport map or generator can be explicitly expressed by using the gradient of the optimal discriminator.

Definition 3. *Let the probability measures μ and ν be defined on discrete sets, $\mathcal{I}$ and $\mathcal{J}$ denote the two disjoint sets of indices. Suppose $\hat{X} = \{x_i\}_{i \in \mathcal{I}}$ and $\hat{Y} = \{y_j\}_{j \in \mathcal{J}}$ are discrete subsets of $\mathbb{R}^d$, and the cost function is defined by $c_{ij} = c(x_i, y_j)$, where $c_{ij} \geqslant 0$ are positive real numbers. Further suppose the source measure $\mu(x_i) = \mu_i$ and the target measure $\nu(y_j) = \nu_j$. A transport plan ρ is a real function that takes values on $\{(x_i, y_j) | \forall x_i \in \hat{\mathcal{X}}, y_j \in \hat{\mathcal{Y}}\}$ such that $\rho_{ij} = \rho(x_i, y_j) \geqslant 0$, $\sum_{i \in \mathcal{I}} \rho_{ij} = \nu_j$ and $\sum_{j \in \mathcal{J}} \rho_{ij} = \mu_i$.*

The total transportation cost $\mathcal{C}$ can be written as:

$$\mathcal{C} = \sum_{i \in \mathcal{I}} \sum_{j \in \mathcal{J}} c_{ij} \rho_{ij} \tag{3}$$

We then calculate the cost of the transportation map from ζ to $\mathcal{Z}$, the one from $\mathcal{Z}$ to Σ, and the one from ζ to Σ. We can first define the measures of them analogously, as follows.

Definition 4. *Let the probability measure μ and ν be defined on discrete sets, $\mathcal{I}_1$ and $\mathcal{J}_1$ denote the two disjoint sets of indices. Suppose $\hat{X}_1 = \{x_i\}_{i \in \mathcal{I}_1}$ and $\hat{Y}_1 = \{y_j\}_{j \in \mathcal{J}_1}$ are discrete subsets of $\mathbb{R}^d$, and the cost function used to map the transportation from random noise to the latent representation is defined by $c_{ij}^1 = c_1(x_i, y_j)$, where $c_{ij}^1 \geqslant 0$ are positive real numbers. Further suppose the source measure $\mu(x_i) = \mu_i$, and the target measure $\nu(y_i) = \nu_j$. A transport plan ρ_1 is a real function that takes values on $\{(x_i, y_j) | \forall x_i \in \hat{\mathcal{X}}_1, y_j \in \hat{\mathcal{Y}}_1\}$ such that $\rho_{ij}^1 = \rho_1(x_i, y_j) \geqslant 0$, $\sum_{i \in \mathcal{I}} \rho_{ij}^1 = \nu_j$ and $\sum_{j \in \mathcal{J}} \rho_{ij}^1 = \mu_i$. This is analogous to c_{ij}^2 and ρ_{ij}^2 defined by the transportation map from the latent representation to the data distribution, and c_{ij}^3 and ρ_{ij}^3 defined by the transportation map from random noise to the data distribution.*

As we know, the latent representation is a low-dimensional mapping related to the data distribution, thus we can assume that there is a linear correlation between c_{ij}^1, c_{ij}^2, and c_{ij}^3, analogously for ρ_{ij}^1, ρ_{ij}^2, and ρ_{ij}^3. Thus, we can say that:

$$\begin{cases} c_{ij}^3 = k_1 c_{ij}^1 = k_2 c_{ij}^2 \\ \rho_{ij}^3 = \bar{k}_1 \rho_{ij}^1 = \bar{k}_2 \rho_{ij}^2 \end{cases} \tag{4}$$

where k_1, k_2, $\bar{k_1}$ and $\bar{k_2}$ are positive numbers to estimate the relationship between them. The total transportation cost from random noise to the data distribution can also be the sum of two components approximately, the transportation cost from random noise to the latent representation and the one from the latent representation to the data distribution. This means $k_1 > 1$ and $k_2 > 1$. Besides, the transportation cost from random noise to the latent representation is significantly lower than the one from the latent representation to the

data distribution since the former has a smaller dimension. Thus, we can say that $k_1 > k_2$. This indicates that $k_1 > k_2 > 1$ and $\bar{k_1} > \bar{k_2} > 1$. Therefore, the total transportation cost of the direct mapping defined as $\mathcal{C}_3$ can be written as:

$$\begin{aligned}\mathcal{C}_3 &= \sum_{i\in\mathcal{I}_3}\sum_{j\in\mathcal{J}_3} c_{ij}^3\rho_{ij}^3 \\ &= \sum_{i\in\mathcal{I}_1}\sum_{j\in\mathcal{J}_1} k_1 c_{ij}^1 \bar{k}_1 \rho_{ij}^1 \\ &= k_1\bar{k}_1 \sum_{i\in\mathcal{I}_1}\sum_{j\in\mathcal{J}_1} c_{ij}^1\rho_{ij}^1 \\ &= k_1\bar{k}_1\mathcal{C}_1 \\ &= k_2\bar{k}_2\mathcal{C}_2\end{aligned} \tag{5}$$

That is to say, the total transportation cost of the direct mapping from random noise to the data distribution is always larger than the one from random noise to the latent representation and the one from the latent representation to the data distribution, which indicates we can reduce the task complexity by Express Construction. Besides, with the decreasing of the task complex, the stability of the model will be further improved by searching for more transportation maps with less transportation cost.

4. Materials and Methods

4.1. Datasets

The proposed method is first evaluated on three small datasets: the prevalent MNIST [17], the Stacked MNIST [18], and the toy dataset [18]. They are well-known datasets that can be used to evaluate the performance of combating the mode collapse problem. The CIFAR-10 dataset [19] will be used to evaluate the performance both quantitatively and qualitatively. Finally, the results on two large-scale datasets named CelebA and CelebA-HQ are described [20,21].

The MNIST of handwritten digits is a subset of a larger set available from NIST, which provides 70,000 examples in total, with 10,000 of them left out for testing. The digits have been size-normalized and centered in a fixed-sized image. Besides, the Stacked MNIST dataset is a dataset in which each image consists of three randomly selected MNIST images that are stacked into a three channels image in RGB that has $10 \times 10 \times 10 = 1000$ modes.

Since the toy dataset consists of explicit distributions and known modes, and the quality of the generated sample can be accurately measured, following [7,18,22–24], we use 2D-ring and 2D-grid to evaluate contributions. Both 2D-ring and 2D-grid are datasets that have a mixture of 25 two-dimensional spherical Gaussians, which can be used to calculate the mode coverage of generating models.

CIFAR-10 is a publicly accessible dataset that contains 50,000 natural images that have been the most widely used for image classification studies and to test the performance of generative models. All images in this dataset are in color with an image size of 32×32 pixels.

CelebFaces Attributes Dataset (CelebA) is a large-scale face attributes dataset with more than 200K celebrity images, each with 40 attribute annotations. The images in this dataset cover large pose variations and background clutter which indicates CelebA has large diversities, large quantities, and rich annotations. Furthermore, another dataset with a higher resolution, namely the CelebA-HQ dataset with the image size of 256×256, is also used to show the performance of the proposed method.

4.2. Method

Express Construction is lightweight method to generate visual images with rich diversity. No extra hyper-parameters have been used in the proposed method, which means Express Construction can be used to train any GAN models without extra training

or hyper-parameter optimization. In this section, we introduce the proposed method in detail.

4.2.1. Framework

In contrast to the traditional game involving a single group of the adversary, two groups of adversaries with different scales are trained to achieve Express Construction. Prior to this, an Auto-Encoder is introduced to embed the data distribution into the latent space. We show the proposed method in Figure 1 and describe it as follows.

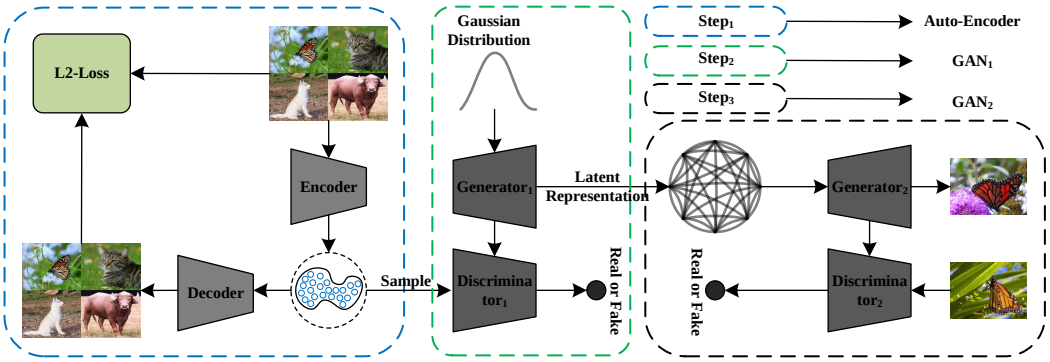


Figure 1. The proposed workflow of Express Construction, all components are parameterized by neural networks. An auto-encoder is trained to map the data manifold into the latent manifold, as shown with the blue dashed box. Then, an adversarial game between a discriminator and a generator with a small scale is used to produce a generated latent representation from random noise, as shown in the green dashed box. Finally, a GAN model with a large scale is trained to construct a transportation map from the result of the previous step to the data distribution, as the black dashed box shows.

4.2.2. Data Embedding with Auto-Encoder

We train an auto-encoder using the L1 Loss to calculate the errors between the data distribution and the reconstructed samples. The encoder is trained to encode the data manifold from the image space to the latent space and map the data distribution to the latent code distribution. Then the decoder decodes the latent code back to the data manifold. In this way, the data distribution is mapped into the latent space, is more abstract but lower-dimensional, and is easier to learn. In particular, training the auto-encoder is equivalent to computing the encoding map f_θ and decoding map g_ξ [15]:

$$(\nu_{gt}, \mathcal{X}) \xrightarrow{f_\theta} (\mu_{gt}, \Omega) \xrightarrow{g_\xi} (\nu_{gt}, \mathcal{X}) \tag{6}$$

where f_θ and g_ξ parameterized by standard convolutional neural networks (CNNs), and $f_\theta : \mathcal{X} \to \Omega$ is an embedding, and pushes forward a probability measure ν_{gt} in $\mathbb{R}^d$ to the latent data distribution μ_{gt}. After training, f_θ is a homeomorphism and g_{xi} is the inverse homeomorphism. This means $f_\theta : \mathcal{X} \to \Omega$ is an embedding, and pushes forward ν_{gt} to the latent data distribution μ_{gt}.

4.2.3. Constructing the Transportation Map from Random Noise to Latent Representation

In practice, we only have the empirical data distribution defined as $\hat{v}_{gt}$, which is pushing forward to be the empirical latent distribution $\hat{\mu}_{gt}$. From the empirical latent distribution, we construct the transportation map from the random noise ζ to the empirical latent representation $\hat{\mu}_{gt}$. Thus, the GAN model is trained to compute the transport map from ζ to $\hat{\mu}_{gt}$ on the manifold.

$$(\zeta, \mathcal{Z}) \xrightarrow{g_\varrho} (\mu, \Omega) \tag{7}$$

where g_ϱ is parameterized by neural networks and μ is a distribution whose support has a similar topology to that of μ_{gt}. We adopt the vanilla GAN model to do this based on the Wasserstein distance. In particular, the Wasserstein distance performs better than Kullback–Leibler (KL), Jensen–Shannon (JS), and total variation (TV) divergence. It correlates with convergence and sample quality, which can serve as a useful metric over probability distribution [25,26]. Following [27], the Wasserstein distance in the dual mode is introduced to evaluate the generated quality, which can be formulated as:

$$\text{Wasserstein}(\mathbb{P}_r, \mathbb{P}_\theta) = \sup_{\|f\|_L \leqslant 1} \mathbb{E}_{x\sim\mathbb{P}_r}[f(x)] - \mathbb{E}_{x\sim\mathbb{P}_\theta}[f(x)] \tag{8}$$

The generator G_1 is used to generate a new latent representation while the discriminator D_1 is used to discriminate if the distribution of the generated samples is the same as that sampled from the empirical latent distribution $\hat{\mu}_{gt}$. Both G_1 and D_1 are built with a small-scale architecture to speed the training up. The training process of is formalized to be a min–max optimization problem, in which the critic solves the following cost function:

$$\mathcal{L}_{D_1} = \mathbb{E}_{z\sim\zeta}[D_1(G_1(z))] - \mathbb{E}_{x\sim\hat{\mu}_{gt}}[D_1(x)] \tag{9}$$

The training cost of this procedure is very small, less than ten minutes. Compared to the AE-OT and its variations, the proposed method is lightweight to produce new distribution close to the empirical latent representation.

4.2.4. Constructing the Transportation Map from Data Latent Representation to Distribution

Instead of training from the random noise, another GAN model with a large scale is trained to construct the transportation map from the generated latent representation to the data distribution, as follows:

$$(\mu, \Omega) \xrightarrow{g_\varsigma} (\nu_{gt}, \mathcal{X}) \tag{10}$$

where g_ς is parameterized by neural networks. The networks in this model are built around two functions: the generator $G_2(\mu)$ maps a sample ν_{gt} from μ, and the discriminator $D_2(x)$ determines if a sample x belongs to the data distribution. They have trained alternately, based on game theory principles. Following the previous step, the Wasserstein distance in the dual mode is introduced to evaluate the sample quality. The training process is also formalized to be a min–max optimization problem. Their cost functions are also followed by the vanilla GAN model, based on the Wasserstein distance. Different models in GANs are then introduced to verify the generalization ability and improve the performance, which will be shown in the experiment.

5. Results

In this section, we conduct extensive experiments to evaluate the performance of Express Construction. Adam optimized with an initial learning rate of 1.00×10^{-4} has been used to train them, and no transform or data augmentation has been utilized in these experiments. The prior noise input to the generator in the first step follows the random Gaussian distribution with 128 sizes. All the experiments are implemented and evaluated with Pytorch on 8× Nvidia Geforce GTX 1080 Ti [28].

The Inception Score (IS) [29] and the Fréchet Inception Distance (FID) [30,31] are used for quantitative evaluation of image quality. The Inception Score is a metric that computes the KL divergence between the conditional class distribution and the marginal class distribution, and Fréchet Inception Distance is a more principled and comprehensive metric that is more consistent with human evaluation in assessing the realism and variation of the generated samples [31,32]. Accuracy in particular is a metric used for classification, which is the ratio between the number of correct predictions and the total number of data points in the dataset. Besides, the number of modes, the percentage of high-quality samples, and the reverse Kullback–Leibler(KL) divergence are used to evaluate the mode coverage.

The number of modes counts the amount of modes captured by samples produced in a generative model. The percentage of high-quality samples measures the proportion of samples generated within three standard deviations of the nearest mode. The reverse KL divergence measures how well generated samples balance among all modes regarding the real distribution.

5.1. MNIST and Stacked MNIST Datasets

We use only 1000 MNIST images to train the networks without any data augmentation, as Figure 2a shows, the contrast of our samples between the foreground and the background is sharp to see. We have also evaluated our performance on the Stacked MNIST dataset, as shown in Figure 2b.

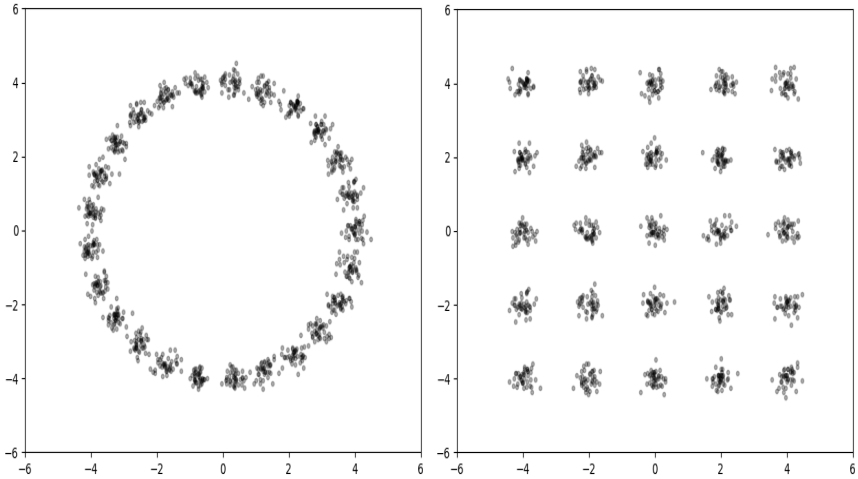
(**a**) MNIST (**b**) Stacked MNIST

Figure 2. MNIST datasets.

The number of observed modes in a generator on this dataset, as well as the KL divergence of the generated mode distributions are measured for evaluation. We generate samples from the generator and each of the three channels in each sample is classified by a pre-trained classifier to determine which of 1000 modes the sample belongs to. Finally, we can fully capture all the modes in the benchmark test, and the empirical KL divergence of them is 0.05 ± 0.008 that is evaluated by 26,000 samples. The results show that the proposed method is effective against the mode collapse problem to capture all the modes in the data distribution.

5.2. Toy Datasets

Under standard benchmark settings in [18], we train the network with 100,000 total samples in toy datasets and a batch size of 100 samples, whether 2D-ring and 2D-grid, as shown in Figure 3. Then, there are ten independent experiments in terms of modes captured, the percentage of high-quality samples, and the reverse KL divergence, and we evaluate them by calculating the average, as shown in Table 1. We achieve the best score of high-quality samples and reverse KL divergence on both datasets. Besides, we achieve a suboptimal result of mode coverage. Significantly, the computational cost of the proposed method is much less than other state-of-the-art methods. On the other hand, Express Construction is much better than the baseline schemes, if the GAN model is used. Therefore, we can say that Express Construction can outperform the baseline schemes and match other state-of-the-art methods, and is less prone to mode collapse.



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

Figure 3. Toy datasets.

Table 1. Experiments on synthetic datasets under standard benchmark settings, in which we can not capture more data modes, but also improve the sample quality. We adopt the vanilla GAN model to train ours$_1$, and the CTGAN to train ours$_2$, respectively. The best score is marked in bold.

	2D-Ring			2D-Grid		
	Modes Max 25	**h-Quality Samples**	**Reverse KL**	**Modes Max 25**	**h-Quality Samples**	**Reverse KL**
GAN [1]	19.7 ± 0.5	95.3 ± 0.2	0.45 ± 0.09	17.3 ± 0.8	94.8 ± 0.7	0.70 ± 0.07
CTGAN [33]	23.8 ± 0.3	98.3 ± 0.1	0.04 ± 0.03	23.5 ± 0.4	98.0 ± 0.2	0.05 ± 0.04
PacGAN [18]	24.7 ± 0.1	96.5 ± 0.3	0.05 ± 0.02	24.6 ± 0.4	94.2 ± 0.4	0.06 ± 0.02
PresGAN [23]	24.7 ± 0.2	97.2 ± 0.3	0.04 ± 0.04	24.7 ± 0.4	94.5 ± 0.2	0.05 ± 0.03
BourGAN [24]	24.8 ± 0.2	97.9 ± 0.1	0.02 ± 0.01	24.9 ± 0.1	95.9 ± 0.2	0.02 ± 0.02
SRGAN [34]	24.8 ± 0.2	97.5 ± 0.2	0.02 ± 0.01	24.7 ± 0.3	98.4 ± 0.3	0.03 ± 0.04
AE-OT [7]	$\mathbf{24.9 \pm 0.1}$	99.8 ± 0.2	$\mathbf{0.01 \pm 0.01}$	$\mathbf{24.9 \pm 0.1}$	99.5 ± 0.5	$\mathbf{0.01 \pm 0.01}$
AE-OT-GAN [15]	24.8 ± 0.2	$\mathbf{99.9 \pm 0.1}$	$\mathbf{0.01 \pm 0.01}$	24.8 ± 0.2	$\mathbf{99.7 \pm 0.3}$	$\mathbf{0.01 \pm 0.01}$
Ours$_1$	22.5 ± 0.5	97.6 ± 0.5	0.08 ± 0.04	22.1 ± 0.7	97.0 ± 0.6	0.10 ± 0.06
Ours$_2$	24.8 ± 0.2	$\mathbf{99.9 \pm 0.1}$	$\mathbf{0.01 \pm 0.01}$	24.8 ± 0.2	$\mathbf{99.7 \pm 0.3}$	$\mathbf{0.01 \pm 0.01}$

5.3. CIFAR-10 Dataset

The base cost function of the GAN used in this experiment is the CTGAN. We first use only 1000 images to train a small neural network, as shown in Figure 4a. The inception score for the result is 5.33 ± 0.10, which is better than the baseline for CTGAN of 5.13 ± 0.12. Then, we have trained another large-scale ResNet [35] on the full training set, as shown in Figure 4b. We achieve an inception score of 8.38 ± 0.15 and an FID of 19.97 ± 0.98 that outperforms the baseline (8.09) by a large margin (+0.29).

We have also tested the performance using different models and compare it with the benchmark, as shown in Table 2. We can achieve a better result no matter what cost function is used in the proposed method.

The same with [33], for the semi-supervised learning approach, we follow the standard training/test split of the dataset but use only 4000 labels in the training. Regular data augmentation with flipping of the images horizontally and randomly translating the images within -2 and 2 pixels is utilized. We report the semi-supervised learning results in Table 3.

Compared to several very competitive methods, Express Construction is able to achieve a state-of-the-art result that outperforms all the GAN-based methods.

(**a**) Generated by a small CNN

(**b**) Generated by a large scale ResNet

Figure 4. CIFAR-10 images generated without supervision by a small and large scale network, respectively.

Table 2. The Inception Score (IS) and the Fréchet Inception Distance (FID) for Express Construction using different cost functions. We can say that it can outperform or match the benchmark. The best score is marked in bold.

GANs	BEGAN	GAN	WGAN	WGAN-GP	CTGAN	WCGAN	SNGAN
IS	5.62	7.01	7.22	7.78	8.09	8.20	8.22
FID	84.0	31.8	30.3	29.5	22.4	20.4	21.7
IS + Express Construction	**6.58**	**7.60**	**7.83**	**8.11**	**8.38**	**8.36**	**8.37**
FID + Express Construction	**55.1**	**29.7**	**28.2**	**27.3**	**19.9**	**20.2**	**20.0**

Table 3. The results of semi-supervised learning methods on the CIFAR-10 dataset. We compare the proposed method with others using the test error. The best score is marked in bold.

GAN-Based	Improved GANs [29]	Improved Semi [36]	CTGAN [33]	Express Construction
	18.63 ± 2.32	16.78 ± 1.80	9.98 ± 0.21	$\mathbf{9.64 \pm 0.25}$

5.4. CelebA

We trained networks using the CelebA dataset for the resolution of 128×128, as shown in Figure 5a. Furthermore, we also test the proposed method on images with high resolution with the CelebA-HQ dataset, as shown in Figure 5b. The cost function is borrowed from BEGAN, and the FID score for them is 14.7 and 6.8, respectively. Finally, Figure 5c is the linear interpolation from a man to a woman, in which the change can be seen clearly.

(**a**) CelebA

(**b**) CelebA-HQ

(**c**) The linear interpolation between given two generated faces

Figure 5. The generated results trained on the CelebA and the CelebA-HQ dataset.

6. Conclusions and Discussion

Generative Adversarial Networks can generate very convincing images, but always fall into the trap of the mode collapse problem. The most important way to tackle this problem is to search for more transportation maps, which has proven very difficult because transportation mapping is always implicit. Searching for relay nodes is a more reliable alternative. Although some works have been proposed to embed the data distribution onto the latent space and select continuous transportation maps, it is costly and potentially reduces the robustness of the model. Thus, how to search transportation maps in a stabler and simpler way is an enormous challenge.

This paper proposes a novel training methodology for GANs to search for more transportation maps with stabler training dynamics and smaller computational costs, named Express Construction. The key idea is to decompose the training of GANs into two phases: an auto-encoder and a small GAN model are trained in the first phase to map the data distribution into the latent space and generate latent representation from random noise, respectively. Then, a large-scale GAN model is trained to generate the distribution closed to the data distribution from the generated latent representation rather than random noise. The proposed method can search for more transportation maps in the latent space and the training dynamics that are stabler than previous works with less computational cost. Besides, no extra hyper-parameters have been used in the proposed method, which indicates that Express Construction can be used to train any GAN models. To the best of our knowledge, Express Construction is the first work that can tackle the mode collapse problem in this way.

Furthermore, Express Construction can achieve a better result by decomposing the generating procedure into more components if we constantly extend the dimension of the latent representation in the latent space. With the increase in the content of the latent representation, the generating performance will be greatly improved. Although Express Construction is lightweight, the computational cost grows multiple times, and the improvement is not corresponding. This is because the computational cost of the proposed two phases will be increased simultaneously with the extension of the latent representation. Therefore, we do not recommend decomposing the generating procedure too many times. Moderate expansion can improve the generating performance and keep the computational cost low.

The proposed method admits the followed extension in the future. It is hard for a single cost function to learn all modes. Instead, multiple discriminators with various cost functions can yield different gradients, which can cover more data modes implicitly. Inspired by this, we attempt to train the GAN models in different phases with different cost functions, thus better tackling the mode collapse problem. There is not any extra computational cost compared to the current work.

Author Contributions: Conceptualization, M.L. (Minghui Liu) and M.L. (Ming Liu); Methodogy, M.L. (Minghui Liu) and J.D.; Software, M.L. (Minghui Liu), M.Y. and P.D.; Validation, J.D. and M.Y.; Formal analysis, M.L. (Minghui Liu) and X.C.; Investigation, X.C. and T.X.; Resources, X.W. and M.L. (Ming Liu); Writing–original draft preparation, M.L. (Minghui Liu) and J.D.; Writing—review and editing, M.Y. and P.D.; Supervision, X.W. and M.L. (Ming Liu); Funding acquisition, X.W. and M.L. (Ming Liu). All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Science and Technology Achievements Transformation Demonstration Project of Sichuan Province of China grant number No. 2018CC0094 and the Fundamental Research Funds for the Central Universities grant number No. ZYGX2019J075.

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

1. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial nets. In Proceedings of the Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, Montreal, QC, Canada, 8–13 December 2014.
2. Aaron van den Oord, N.K.; Kavukcuoglu, K. Pixel Recurrent Neural Networks. In Proceedings of the International Conference on Machine Learning (ICML), New York, NY, USA, 19–24 June 2016; pp. 1747–1756.
3. Van den Oord, A.; Kalchbrenner, N.; Espeholt, L.; Vinyals, O.; Graves, A. Conditional Image Generation with PixelCNN Decoders. In Proceedings of the Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, Barcelona, Spain, 5–10 December 2016.
4. Kingma, D.; Max, W. Auto-Encoding Variational Bayes. In Proceedings of the International Conference on Learning Representations (ICLR), Banff, AB, Canada, 14–16 April 2014.
5. Tolstikhin Ilya, B.O.; Sylvain, G. Wasserstein Auto-Encoders. In Proceedings of the International Conference on Learning Representations(ICLR), Vancouver, BC, Canada, 30 April–3 May 2018.
6. Guo, Y.; An, D.; Qi, X.; Luo, Z.; Yau, S.T.; Gu, X. Mode Collapse and Regularity of Optimal Transportation Maps. *arXiv* **2019**, arXiv:1902.02934.
7. An, D.; Guo, Y.; Lei, N.; Luo, Z.; Yau, S.T.; Gu, X. AE-OT: A New Generative Model Based on Extended Semi-discrete Optimal Transport. In Proceedings of the International Conference on Learning Representations (ICLR), Addis Ababa, Ethiopia, 26–30 April 2020.
8. Nagarajan, V.; Kolter, J.Z. Gradient Descent GAN Optimization Is Locally Stable. In *Advances in Neural Information Processing Systems (NIPS)*; MIT Press: Long Beach, CA, USA, 2017; Volume 30.
9. Tenenbaum, J.B.; Silva, V.; Langford, J.C. A global geometric framework for nonlinear dimensionality reduction. *Science* **2000**, *290*, 2319–2323. [CrossRef] [PubMed]
10. Liou, C.Y.; Huang, J.C.; Yang, W.C. Modeling word perception using the Elman network. *Neurocomputing* **2008**, *71*, 3150–3157. [CrossRef]
11. Hoang, Q.; Nguyen, T.D.; Le, T.; Phung, D. MGAN: Training Generative Adversarial Nets with Multiple Generators. In Proceedings of the International Conference on Learning Representations (ICLR), Vancouver, BC, Canada, 30 April–3 May 2018.
12. Han, C.; Rundo, L.; Murao, K.; Noguchi, T.; Shimahara, Y.; Milacski, Z.Á.; Koshino, S.; Sala, E.; Nakayama, H.; Satoh, S.I. MADGAN: Unsupervised medical anomaly detection GAN using multiple adjacent brain MRI slice reconstruction. *BMC Bioinform.* **2021**, *22*, 31. [CrossRef] [PubMed]
13. Metz, L.; Poole, B.; Pfau, D.; Sohl-Dickstein, J. Unrolled Generative Adversarial Networks. In Proceedings of the International Conference on Learning Representations (ICLR), Toulon, France, 24–26 April 2017.
14. Liu, H.; Guo, Y.; Lei, N.; Shu, Z.; Yau, S.T.; Samaras, D.; Gu, X. Latent Space Optimal Transport for Generative Models. *arXiv* **2018**, arXiv:1809.05964.
15. An, D.; Guo, Y.; Zhang, M.; Qi, X.; Lei, N.; Gu, X. AE-OT-GAN: Training GANs from Data Specific Latent Distribution. In Proceedings of the European Conference on Computer Vision (ECCV), Glasgow, UK, 23–28 August 2020; Springer: Berlin/Heidelberg, Germany, 2020; pp. 548–564.
16. Arjovsky, M.; Bottou, L. Towards Principled Methods for Training Generative Adversarial Networks. In Proceedings of the International Conference on Learning Representations (ICLR), Toulon, France, 24–26 April 2017.
17. LeCun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. Gradient-based learning applied to document recognition. *Proc. IEEE* **1998**, *86*, 2278–2324. [CrossRef]
18. Lin, Z.; Khetan, A.; Fanti, G.; Oh, S. PacGAN: The power of two samples in generative adversarial networks. In *Advances in Neural Information Processing Systems (NIPS)*; MIT Press: Montreal, QC, Canada, 2018; Volume 31.
19. Krizhevsky, A.; Hinton, G.E. *Learning Multiple Layers of Features from Tiny Images*; Technical Report; Citeseer: Princeton, NJ, USA, 2009.
20. Liu, Z.; Luo, P.; Wang, X.; Tang, X. Deep Learning Face Attributes in the Wild. In Proceedings of the International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015.
21. Tero, K.; Timo, A.; Samuli, L.; Jaakko, L. Progressive Growing of GANs for Improved Quality, Stability, and Variation. In Proceedings of the International Conference on Learning Representations (ICLR), Vancouver, BC, Canada, 30 April–3 May 2018.
22. Dumoulin, V.; Belghazi, I.; Poole, B.; Mastropietro, O.; Lamb, A.; Arjovsky, M.; Courville, A. Adversarially Learned Inference. *arXiv* **2016**, arXiv:1606.00704
23. Dieng, A.B.; Ruiz, F.J.; Blei, D.M.; Titsias, M.K. Prescribed Generative Adversarial Networks. *arXiv* **2019**, arXiv:1910.04302.
24. Xiao, C.; Zhong, P.; Zheng, C. BourGAN: Generative Networks with Metric Embeddings. In *Advances in Neural Information Processing Systems (NIPS)*; MIT Press: Montreal, QC, Canada, 2018; Volume 31.
25. Gulrajani, I.; Ahmed, F.; Arjovsky, M.; Dumoulin, V.; Courville, A.C. Improved Training of Wasserstein GANs. In *Advances in Neural Information Processing Systems (NIPS)*; MIT Press: Long Beach, CA, USA, 2017; Volume 30.
26. Salimans, T.; Zhang, H.; Radford, A.; Metaxas, D. Improving GANs Using Optimal Transport. *arXiv* **2018**, arXiv:1803.05573.
27. Martin Arjovsky, S.C.; Bottou, L. Wasserstein GAN. *arXiv* **2017**, arXiv:1701.07875.

28. Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems (NIPS)*; MIT Press: Vancouver, BC, Canada, 2019; Volume 32.
29. Tim, S.; Ian, G.; Wojciech, Z.; Vicki, C.; Alec, R.; Chen, X. Improved Techniques for Training GANs. In *Advances in Neural Information Processing Systems (NIPS)*; MIT Press: Barcelona, Spain, 2016.
30. Dowson, D.; Landau, B. The frechet distance between multvariate normal distributions. *J. Multivar. Anal.* **1982**, *12*, 450–455. [CrossRef]
31. 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 *Advances in Neural Information Processing Systems(NIPS)*; MIT Press: Long Beach, CA, USA, 2017; Volume 30.
32. Zhang, H.; Goodfellow, I.; Metaxas, D.; Odena, A. Self-Attention Generative Adversarial Networks. In Proceedings of the International Conference on Machine Learning (ICLR), Long Beach, CA, USA, 9–15 June 2019; pp. 7354–7363.
33. Wei, X.; Gong, B.; Liu, Z.; Lu, W.; Wang, L. Improving the Improved Training of Wasserstein GANs: A Consistency Term and Its Dual Effect. In Proceedings of the International Conference on Machine Learning (ICLR), Vancouver, BC, Canada, 30 April–3 May 2018.
34. Liu, K.; Tang, W.; Zhou, F.; Qiu, G. Spectral Regularization for Combating Mode Collapse in GANs. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea, 27 October–2 November 2019; pp. 6382–6390.
35. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
36. Kumar, A.; Sattigeri, P.; Fletcher, P.T. Improved Semi-supervised Learning with GANs using Manifold Invariances. *arXiv* **2017**, arXiv:1705.08850.