A High-Quality Robust Diffusion Framework for Corrupted Dataset

Quan Dao[1,2†‡], Binh Ta[1†], Tung Pham[1], and Anh Tran[1]

[1] VinAI Research
v.{quandm7,binhth5,tungph4,anhtt152}@vinai.io
[2] Rutgers University
quan.dao@rutgers.edu

Abstract. Developing image-generative models, which are robust to outliers in the training process, has recently drawn attention from the research community. Due to the ease of integrating unbalanced optimal transport (UOT) into adversarial framework, existing works focus mainly on developing robust frameworks for generative adversarial model (GAN). Meanwhile, diffusion models have recently dominated GAN in various tasks and datasets. However, according to our knowledge, none of them are robust to corrupted datasets. Motivated by DDGAN, our work introduces the first robust-to-outlier diffusion. We suggest replacing the UOT-based generative model for GAN in DDGAN to learn the backward diffusion process. Additionally, we demonstrate that the Lipschitz property of divergence in our framework contributes to more stable training convergence. Remarkably, our method not only exhibits robustness to corrupted datasets but also achieves superior performance on clean datasets.

Keywords: Diffusion Model · Unbalanced Optimal Transport · Robustness Generation · OT-based Generative Model

1 Introduction

In recent years, generative models have seen remarkable advancements. These models have demonstrated the ability to generate pieces of writing, create stunning images, and even produce realistic videos in response to arbitrary queries. However, training datasets often originate from diverse sources, inevitably containing outliers resulting from various factors such as human error or machine inaccuracies. These outliers could significantly impede the performance of models; for instance, a generative model affected by outliers may produce undesired samples. In this study, we focus on a specific scenario where the training dataset for generative model is corrupted by outliers.

[‡] Work done during VinAI internship.

[†] Equal contribution.

The aforementioned scenario has been previously explored in the works [4,56], primarily focusing on Generative Adversarial Networks (GANs). By leveraging unbalanced optimal transport (UOT), [4] proposed RobustGAN to enhance model robustness by using the third weight network to assign less attention to outliers and focus more on clean data. However, this approach not only requires additional training time and resources but also suffers from training instability due to the optimization of three networks, impairing the model's ability to create realistic images. Recently, [39] introduced OTM, a novel type of generative model known as the OT-based generative model, where the optimal transport map itself serves as a generative model. Building upon this work, [6] proposed UOTM framework which replaces the UOT formulation in the OT-based generative model. UOTM demonstrates strong performance on clean datasets, thereby bringing the OT-based generative model on par with other types of generative models such as diffusion and GANs in terms of quality. However, it is worth noting that UOTM only conducts robustness experiments on small-scale datasets with simplified settings, which may not accurately reflect real-world scenarios.

In addition to GANs, recent diffusion models [15, 44, 47, 48] have experienced rapid growth due to their capability to outperform GANs in generating highly realistic images. These models offer adaptability in handling a wide range of conditional inputs, including semantic maps, text, and images, as highlighted in the works of [31, 38, 40, 52]. Despite these immense potentials, diffusion models face a significant weakness: slow sampling speed, as they require extremely large models with thousands of steps to slowly refine an image of white noise into a high-quality picture. This limitation impedes their widespread adoption, contrasting them with GANs. Hence, the combination of GANs and diffusion models, introduced in Denoising Diffusion GAN (DDGAN) [55], has effectively addressed the challenge of modeling complex multimodal distributions, particularly when dealing with large step sizes, through the utilization of GANs. This innovative approach has led to a significant reduction in the number of denoising steps required, typically just a few (e.g., 2 or 4). On the other hand, robust generation is a critical issue frequently encountered in real-world scenarios. While this problem has been extensively studied in recent years, particularly in the context of GANs, it is evident that GANs still lag behind diffusion models in terms of image synthesis quality. Consequently, there is a growing consensus that diffusion models are poised to replace GANs as the leading approach in generative modeling. Given this shift in focus, it becomes imperative to address the question of how to train robust diffusion models that can effectively handle real-world datasets. To date, the development of robust diffusion models tailored for datasets containing a mixture of clean and outlier data points remains largely unexplored. Our work aims to fill this gap by proposing a robust diffusion framework capable of harnessing the high-quality synthesis capabilities of diffusion models while ensuring robustness throughout the generation process.

To address the challenge of producing a high-quality and fast sampling diffusion model in the presence of corrupted data, a straightforward solution might seem to be a combination of DDGAN and UOT, leveraging the strengths of both approaches. However, our work demonstrates that a simple combination of these techniques does not effectively solve the problem. Firstly, we demonstrate that DDGAN utilizes optimal transport (OT) to minimize the probability distance between fake and true distributions, whereas UOT learn to minimize the mapping between source and target distributions. Consequently, GAN and UOT have distinct objectives, making their direct combination challenging. Integrating UOTM into the GAN framework requires additional weight networks [4], leading to poor convergence. In contrast, an OT-based generative framework [39] can seamlessly replace the UOT loss, as both share the same optimization objective. Motivated by this insight, we propose replacing the GAN process in DDGAN with an OT-based generative model to learn the backward diffusion process $q(x_{t-1}|x_t)$, facilitating the integration of the UOT loss. However, we discover that simply modeling $q(x_{t-1}|x_t)$ by UOT is ineffective because large t makes it harder for UOT to distinguish between outliers and clean samples from $p(x_{t-1})$. To address this challenge, we propose learning the distribution $q(x_0|x_t)$ instead, as the UOT loss can more effectively filter out outliers from $q(x_0)$. Additionally, we highlight the effectiveness of Lipschitz Ψ in stabilizing the training of the proposed framework. We summarize our contributions as follows:

- **Robust Diffusion UOT Framework**: We propose a novel approach to integrate UOT into the DDGAN framework by replacing the GAN process with an OT-based generative model. To address the challenge of distinguishing outliers from clean samples as diffusion steps increase, we propose to learn the distribution $p(x_0|x_t)$ instead of $q(x_{t-1}|x_t)$, leveraging the effectiveness of UOT in filtering outliers from the clean distribution $q(x_0)$.
- **Lipschitz Ψ makes stable training**: We emphasize the importance of Lipschitz Ψ in stabilizing the training process of our proposed framework, contributing to its overall effectiveness and stability.
- **Fast, High-fidelity, and Robust Image Generation**: Our proposed model exhibits superior performance compared to DDGAN and UOTM on clean datasets. Moreover, our framework demonstrates enhanced robustness, achieving a lower FID compared to other methods designed for robustness.

2 Related work

In this section, we summarise the related works about unbalanced optimal transport (UOT) in generative models and diffusion models.

UOT in generative models: [3] proposed WGAN which showed the benefits of applying OT in GAN, which minimizes the Wasserstein distance between real and generated distribution. Indeed, OT theory has been the subject of extensive research over an extended period [2,8,17,34,35,51]. This has led to techniques aimed at enhancing the efficiency of OT within GAN models [42,43], all of which utilize Wasserstein distance. Among the variants of OT, Unbalanced OT (UOT) has the potential to make a model more robust to training outliers [12]. Recent works [4, 56] proposed to integrate the UOT loss into GAN framework. However, these works need three distinct neural networks which leads to poor convergence and low-quality image synthesis. [39] proposed an OT-based generative model that optimal transport (OT) map itself can be used as a generative model. Recently, UOTM [6] extended OT-based generative model to UOT-based generative model by replacing OT formula with UOT formula. Though UOTM works well for clean datasets, its robustness experiments are only limited to low-resolution datasets. In this work, we show that our framework by extending the UOT-based generative model for diffusion framework achieves SoTA FID score at both clean and corrupted datasets.

Diffusion models: Diffusion models outperform state-of-the-art GANs in terms of high-fidelity synthesized images on various datasets [10, 41]. Furthermore, diffusion models also possess superior mode coverage [16,23,46], and offer adaptability in handling a wide range of conditional inputs including semantic maps, text, and images [31, 38, 52]. This flexibility has led to their application in various areas, such as text-to-image generation, image-to-image translation, image inpainting, image restoration, and more [28,37,40,41]. Nonetheless, their real-life application was shadowed by their slow sampling speed. DDPM [15] requires a thousand sampling steps to obtain the high-fidelity image, resulting in long-time sampling. Although several techniques have been designed to reduce inference time [30, 45, 57], primarily through reduction of sampling steps, they still need more than 10 NFEs to generate images, roughly 10 times slower than GANs. Recently, DDGAN [55] utilized GAN to tackle the challenge of modeling complex multimodal distributions caused by large step sizes. This model needs much fewer steps (e.g. 2 or 4) to generate an image.

3 Background

3.1 Unbalanced Optimal Transport

In this section, we provide some background on optimal transport (OT), its unbalanced formulation (UOT), and its applications.

Optimal Transport: Let μ and ν be two probability measures in the set of probability measures $\mathcal{P}(\mathcal{X})$ for space $\mathcal{X}$, the OT distance between μ and ν is defined as:

$$\mathsf{OT}(\mu, \nu) = \min_{\pi \in \Pi(\mu,\nu)} \int c(x, y) d\pi(x, y), \tag{1}$$

where $c : \mathcal{X} \times \mathcal{X} \to [0, \infty)$ is a cost function, $\Pi(\mu, \nu)$ is the set of joint probability measures on $\mathcal{X} \times \mathcal{X}$ which has μ and ν as marginal probability measures. The dual form of OT is:

$$\mathsf{OT}(\mu, \nu) = \sup_{u(x)+v(y) \leq c(x,y)} \int_{\mathcal{X}} u(x) d\mu(x) + \int_{\mathcal{X}} v(y) d\nu(y). \tag{2}$$

Denote $v^c(x) = \inf_{y \in \mathcal{X}} \left\{ c(x, y) - v(y) \right\}$ to be the c-transform of $v(y)$, then the dual formulation of OT could be written in the following form:

$$\mathsf{OT}(\mu, \nu) = \sup_{v} \int_{\mathcal{X}} v^c(x) d\mu(x) + \int_{\mathcal{X}} v(y) d\nu(y).$$

Unbalanced Optimal Transport: A more generalized version of OT introduced by [5] is Unbalanced Optimal Transport (UOT) formulated as follows:

$$\mathsf{UOT}(\mu,\nu) = \min_{\pi\in\mathcal{M}(\mathcal{X}\times\mathcal{X})} \int \tau c(x,y) d\pi(x,y) + \mathsf{D}_{\Psi_1}(\pi_1\|\mu) + \mathsf{D}_{\Psi_2}(\pi_2\|\nu), \qquad (3)$$

where $\mathcal{M}(\mathcal{X}\times\mathcal{X})$ denotes the set of joint non-negative measures on $\mathcal{X}\times\mathcal{X}$; π is an element of $\mathcal{M}(\mathcal{X}\times\mathcal{X})$, its marginal measures corresponding to μ and ν are π_1 and π_2, respectively; the D_{Ψ_i} are often set as the Csiszár-divergence, i.e., Kullback-Leibler divergence, χ^2 divergence, τ is a hyper-parameter acting as the weight for the cost function. In contrast to OT, the UOT does not require hard constraints on the marginal distributions, thus allowing more flexibility to adapt to different situations. Similar to the OT, solving the UOT again could be done through its dual form [5,12,49].

$$\mathsf{UOT}(\mu,\nu) = \sup_{u(x)+v(y)\leq\tau c(x,y)} \int_{\mathcal{X}} -\Psi_1^*(-u(x))d\mu(x) + \int_{\mathcal{X}} -\Psi_2^*(-v(y))d\nu(y), \qquad (4)$$

where $u, v \in \mathcal{C}(\mathcal{X})$ in which $\mathcal{C}$ denotes a set of continuous functions over its domain; Ψ_1^* and Ψ_2^* are the convex conjugate functions of Ψ_1 and Ψ_2, respectively. If both function Ψ_1^* and Ψ_2^* are non-decreasing and differentiable, we could next remove the condition $u(x)+v(y)\leq\tau c(x,y)$ by the c-transform for function v to obtain the semi-dual UOT form [49], v is 1-Lipschitz:

$$\mathsf{UOT}(\mu,\nu) = \sup_{||v||_L\leq 1} \int_{\mathcal{X}} -\Psi_1^*\big(-v^c(x)\big)d\mu(x) + \int_{\mathcal{X}} -\Psi_2^*\big(-v(y)\big)d\nu(y). \qquad (5)$$

Follow the definition of c-transform, UOTM [6] write $v^c(x) = \inf_{\hat{x}\in\mathcal{X}} \tau c(x,\hat{x}) - v(\hat{x})$ where both optimal value of generated data $\hat{x}$ and potential function v are unknown. Therefore, UOTM finds the function v through learning a parameterized potential network D_ϕ and optimizing a parameterized generator $G_\theta : \mathcal{X} \to \mathcal{X}$ as mapping from input x to the optimal value of $\hat{x}$. Therefore, Eq. (5) can be written as follows:

$$\begin{aligned}\mathsf{UOT}(\mu,\nu) =& \sup_{D_\phi}\Big[\int_{\mathcal{X}} \Psi_1^*\Big(-\Big[\tau c\big(x,G_\theta(x)\big) - D_\phi\big(G_\theta(x)\big)\Big]\Big)d\mu(x) \\ &+ \int_{\mathcal{X}} \Psi_2^*\big(-D_\phi(y)\big)d\nu(y)\Big] \qquad (6)\\ =& \inf_{D_\phi}\Big[\int_{\mathcal{X}} \Psi_1^*\Big(-\inf_{G_\theta}\Big[\tau c\big(x,G_\theta(x)\big) - D_\phi\big(G_\theta(x)\big)\Big]\Big)d\mu(x) \\ &+ \int_{\mathcal{X}} \Psi_2^*\big(-D_\phi(y)\big)d\nu(y)\Big]. \qquad (7)\end{aligned}$$

3.2 Diffusion Models

Diffusion models that rely on the diffusion process often take empirically thousand steps to diffuse the original data to become a neat approximation of Gaussian noise. Let's use x_0 to denote the true data, and x_t denotes that datum after t steps of rescaling data and adding Gaussian noise. The probability distributions of x_t conditioned on x_{t-1} and x_0 has the form

$$q(x_t|x_{t-1}) = \mathcal{N}(\sqrt{1-\beta_t}x_{t-1}, \beta_t\mathbf{I}) \tag{8}$$

$$q(x_t|x_0) = \mathcal{N}(x_t; \sqrt{\overline{\alpha}_t}x_0, (1-\overline{\alpha}_t)\mathbf{I}) \tag{9}$$

where $\alpha_t = 1 - \beta_t$, $\overline{\alpha}_t = \prod_{s=1}^{t} \alpha_s$, and $\beta_t \in (0, 1)$. Since the forward process introduces relatively minor noise each step, we can approximate reverse probability $p(x_{t-1}|x_t)$ using Gaussian probability $q(x_{t-1}|x_t, x_0)$, which could be learned through a parameterized function $p_\theta(x_{t-1}|x_t)$. Following [15], $p_\theta(x_{t-1}|x_t)$ is commonly parameterized as:

$$p_\theta(x_{t-1} \mid x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \sigma_t^2\mathbf{I}), \tag{10}$$

where $\mu_\theta(x_t, t)$ and σ_t^2 represent the mean and variance of parameterized denoising model, respectively. The learning objective is to minimize the Kullback-Leibler (KL) divergence between true denoising distribution $q(x_{t-1}|x_t)$ and denoising distribution parameterized by $p_\theta(x_{t-1}|x_t)$.

Unlike traditional methods, DDGAN [55] allows for larger denoising step sizes to speed up the sampling process by incorporating generative adversarial networks (GANs). DDGAN introduces a discriminator, denoted as D_ϕ, and optimizes both the generator and discriminator in an adversarial training fashion. The objective of DDGAN can be expressed as follows:

$$\begin{aligned}\min_{\phi} \max_{\theta} \sum_{t \geq 1} \mathbb{E}_{q(\mathbf{x}_t)} \Big\{ & \mathbb{E}_{q(\mathbf{x}_{t-1}|\mathbf{x}_t)} \Big[-\log \big(D_\phi(\mathbf{x}_{t-1}, \mathbf{x}_t, t) \big) \Big] \\ & + \mathbb{E}_{p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)} \Big[\log \big(D_\phi(\mathbf{x}_{t-1}, \mathbf{x}_t, t) \big) \Big] \Big\}\end{aligned} \tag{11}$$

In Eq. (11), conditional generator $p_\theta(x_{t-1}|x_t)$ generates fake samples. Due to large step sizes, the distribution $q(x_{t-1}|x_t)$ is no longer Gaussian. DDGAN models this complex multimodal distribution by using a generator $G_\theta(x_t, z, t)$, where z is a D-dimensional latent variable drawn from a standard Gaussian distribution $\mathcal{N}(0, \mathbf{I})$. Specifically, DDGAN first generates an clean sample x_0' through the generator $G_\theta(x_t, z, t)$ and obtains the perturbed sample x_{t-1}' using $q(x_{t-1}|x_t, x_0')$. Simultaneously, the discriminator evaluates both real pairs $D_\phi(x_{t-1}, x_t, t)$ and fake pairs $D_\phi(x_{t-1}', x_t, t)$ to guide the training process.

4 Method

Recent works [4,6] on robust generative models replace OT with UOT in adversarial framework. However, GANs are widely known for training instability and mode collapse [24]. By combining diffusion process and GAN models, Denoising Diffusion GAN (DDGAN) [55] successfully mitigates these limitations. While

GAN uses OT distance to minimize the moving cost between real and fake distributions, UOT formulation minimizes the moving cost from source to target distributions. Therefore, it is hard to directly apply UOT into DDGAN framework. In the Sec. 4.1, motivated by OT-based generative [6, 39], we model backward diffusion process $p(x_{t-1}|x_t)$ by UOT-based generative model for robust-to-outlier image generation. However, naively modelling $p(x_{t-1}|x_t)$ leads to high FID since diffusion noising process reduces the difference between outlier and clean data. Instead, we model $p(x_0|x_t)$ by a UOT-based generative model to easily eliminate outliers. Sec. 4.2 presents the importance of Lipschizt property of Ψ and how to design the potential network D_ϕ, generator network G_θ.

4.1 Robust-to-Outlier Diffusion Framework

DDGAN matches the conditional GAN generator $p_\theta(x_{t-1}|x_t)$ and $q(x_{t-1}|x_t)$ using an adversarial framework that minimizes OT loss per denoising step:

$$\min_\theta \sum_{t\geq 1} \mathbb{E}_{q(x_t)} \mathsf{OT}\left(q\left(x_{t-1} \mid x_t\right) \| p_\theta\left(x_{t-1} \mid x_t\right)\right) \tag{12}$$

where $q(x_{t-1}|x_t)$ is ground-truth conditional distribution with x_{t-1} sampling from Eq. (9) and x_t sampling from Eq. (8). The fake conditional pair $(\hat{x}_{t-1}, x_t) \sim p_\theta(x_{t-1}|x_t)$ is obtained using ground truth x_t and $\hat{x}_{t-1} \sim q(x_{t-1}|x_t, \hat{x}_0)$ with $\hat{x}_0 = G_\theta(x_t, z, t)$ $(z \sim \mathcal{N}(0, \mathbb{I}))$. Noted that: In DDGAN, **OT cost serves as the loss** to minimize that distance between true distribution $q(x_{t-1}|x_t)$ and fake distribution $p_\theta(x_{t-1}|x_t)$. For robustness problem, we cannot directly apply UOT formulation into GAN-based architecture since UOT does not measure the distance between true and fake distributions. To apply UOT in GAN, RobustGAN [4] needs additional network W to weight the outliers, which leads to training instability due to optimization of three networks.

Motivated from [6, 39], instead of minimizing OT cost between $q(x_{t-1}|x_t)$ and $p_\theta(x_{t-1}|x_t)$, our framework uses **optimal transport map as a generative model itself**, which is an OT-based generative model [6, 39]. To enable robustness property, we aim to learn a UOT mapping from marginal distribution $q(x_t)$ to backward diffusion process $q(x_{t-1}|x_t)$.

$$\sum_{t\geq 1} \mathsf{UOT}\left(q(x_t), q\left(x_{t-1}|x_t\right)\right) \tag{13}$$

However, due to diffusion process, the robustness property of generative model trained by Eq. (13) is not guaranteed. In Eq. (3), if τ is too small, UOT formulation becomes an OT formulation which penalizes the marginal constraints and ignores the outlier filtering. In contrast, when τ is too large, UOT formulation focus more to outlier filtering and ignores the marginal constraints. In case, the outlier and clean distributions are close to each other, τ should be increased for robustness guarantee. By Proposition 1 (proof in Appendix 8), the outlier and clean noisy samples at time t become close to each other as t increases and **the τ should also increase as t increases**. It is hard to cast out the outlier among x_t since **choosing different τ for each step t costs a huge amount of time and resource**. Furthermore, when the outlier and clean noisy samples for large t are too similar, **large τ could accidentally remove the low-density modality of clean distribution and cannot eliminate the outlier samples**.

Proposition 1. *Denote P^c and P^o be clean and outlier probability measures. Let P_t be the probability measure that $x_t \sim P_t$ is obtained from $x_0 \sim P$ by a forward diffusion. Wasserstein* ***distance $W(P_t^c, P_t^o)$ decreases as t increases.***

To solve this problem, we use UOT to map from marginal distribution $q(x_t)$ to backward diffusion $q(x_0|x_t)$, shown in Eq. (14). The backward diffusion $q(x_{t-1}|x_t)$ is intractable [15] and it could be written as $q(x_{t-1}|x_t) = \sum_{x_0} q(x_{t-1}|x_t, x_0)q(x_0|x_t)$. From this observation, we formulate the following loss for our framework:

$$\sum_{t\geq 1} \mathsf{UOT}\left(q\left(x_t\right), q\left(x_0 \mid x_t\right)\right) \tag{14}$$

There are two motivating reasons for using Eq. (14). Firstly, since x_0 is zero-noised, the distance between outlier and inlier x_0 is large and UOT formulation could effectively remove the outliers. This formula helps us avoid the robust ill-posed problem stated by Proposition 1. Secondly, we notice that $q(x_{t-1}|x_t, x_0)$ [15] is tractable and could be easily sampled due to its Gaussian form. Applying the semi-dual UOT Eq. (7) in the training objective Eq. (14), we can obtain:

$$\begin{aligned}\mathsf{UOT}(q\left(x_t\right), q\left(x_0 \mid x_t\right)) = \min_{D_\phi} \Bigg[& \Psi_1^*\Big(- \min_{G_\theta} \Big[\tau c\big(x_t, \hat{x}_0\big) - D_\phi\big(\hat{x}_0, x_t, t\big)\Big]\Big) \\ & + \Psi_2^*\big(- D_\phi(x_0, x_t, t)\big)\Bigg],\end{aligned} \tag{15}$$

where $\hat{x}_0 = G_\theta(x_t, t)$.

4.2 Analysis of Semi-Dual UOT formulation

In this section, we analyze the importance of choosing Ψ in Eq. (15), the design space of potential network D_ϕ and G_θ.

Lipschitz property of Ψ: UOTM [6] favour the conventional Csiszár-divergence Ψ like KL or χ^2. However, in Sec. 5.3, we show that the function, whose convex conjugate is Softplus, performs better than these conventional divergences. As [1] states that the Lipschitz loss function results in better performance, we hypothesize that Lipschitz continuity property of Softplus helps the training process more effective while convex conjugate of KL and χ^2 are not Lipschitz (see Appendix 9 for proof of Lipschitz property).

Design space of generator function G_θ: Motivated from [55], we also inject latent variable $z \sim \mathcal{N}(0, I)$ as input to G_θ along with x_t and t. There are two reasons for this choice. Firstly, the latent variable z helps the generator mimic stochastic behavior. According to [55], without latent z, the denoising generative model becomes a unimodal distribution, making the sample quality significantly worse. The second reason is that z can be used as style information as in StyleGAN architecture [21]. Motivated from StyleGAN, DDGAN generator network [55] also uses style modulation layer and AdaIn to inject style information from z into each feature network. As a result, DDGAN inherits the sophisticated architecture of StyleGAN for high-fidelity image synthesis. We adopt a similar architecture design of generator G_θ from DDGAN [55].

Design space of potential function D_ϕ: Through experiment, we discover that using x_{t-1} (instead of x_0 in Eq. (15)) in potential network D_ϕ in place for x_0 achieves better FID score. In sampling process, given x_t, we predict $\hat{x}_0 = G_\theta(x_t, t, z)$ then draw $x_{t-1} \sim q(x_{t-1}|x_t, \hat{x}_0)$, consequently. The sampling process not only depends on $G_\theta(x_t, t, z)$ but also $q(x_{t-1}|x_t, \hat{x}_0)$. Therefore, in training framework, we should explicitly use x_{t-1} from $q(x_{t-1}|x_t, \hat{x}_0)$ as input of potential network to better support the sampling process. Relying on the reason, we propose the modified UOT loss replacing Eq. (15):

$$\begin{aligned}\mathsf{UOT}(q\left(x_t\right), q\left(x_0 \mid x_t\right)) = \min_{D_\phi} \Bigg[& \Psi_1^*\Big(- \min_{G_\theta} \Big[\tau c\big(x_t, \hat{x}_0\big) - D_\phi\big(\hat{x}_{t-1}, x_t, t\big)\Big]\Big) \\ & + \Psi_2^*\big(- D_\phi(x_{t-1}, x_t, t)\big)\Bigg], \end{aligned} \tag{16}$$

where $x_{t-1} \sim q(.|x_t, x_0)$.

In summary, we present our framework ***Robust Diffusion Unbalanced Optimal Transport (RDUOT)*** in Algorithm 1. In the default setting on clean dataset and outlier robustness, we apply semi-dual UOT to all diffusion steps and use the same cost functions $\mathbf{L}_2$: $c(x, y) = \tau||x - y||_2^2$ as UOTM.

5 Experiment

In this section we firstly show the robustness of our model RDUOT to various corrupted datasets. We then show that RDUOT also possesses high-fidelity generation and fast training convergence properties on clean datasets. Finally, we conduct ablation studies to show the importance of choosing Ψ, and to verify the design of our framework in Sec. 4. Details of all experiments and evaluations can be found in Appendix 7.

5.1 Robustness to corrupted datasets

In this section, we conducted experiments on various datasets perturbed with diverse outlier types, mirroring real-world applications to validate its robustness in handling corrupted datasets. Since the resolution of clean and outlier datasets might be different, we rescaled the clean and outlier datasets to the same resolution, with CI+MI at 32×32 and the other four datasets (CE+FT, CE+MT,

```
Algorithm 1: Robust Diffusion Unbalanced Optimal Transport
  Input: The data distribution $p_{data}$. Non-decreasing, differentiable, a function
    pair $(\Psi_1^*, \Psi_2^*)$. Generator network $G_\theta$ and the potential network $D_\phi$. Total
    training iteration number $K$. Batch size $B$.
  for $k = 0, 1, 2, \ldots, K$ do
      Sample $x_0 \sim p_{\mathrm{data}}, z \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d), t \sim [1 : T]$.
      Sample $x_t \sim p(\cdot|x_0), \hat{x}_0 = G_\theta(x_t, z, t), \hat{x}_{t-1} \sim p(\cdot|\hat{x}_0, x_t), x_{t-1} \sim p(\cdot|x_0, x_t)$.
          $\mathcal{L}_D = \frac{1}{B}\Psi_1^*\big(-c(x_t, \hat{x}_0) + D_\phi(\hat{x}_{t-1}, x_t, t)\big) + \frac{1}{B}\Psi_2^*\big(-D_\phi(x_{t-1}, x_t, t)\big).$
      Update $\phi$ to minimize the loss $\mathcal{L}_D$.
          $\mathcal{L}_G = \frac{1}{B}\big(c(x_t, \hat{x}_0) - D_\phi(\hat{x}_{t-1}, x_t, t)\big).$
      Update $\theta$ to minimize the loss $\mathcal{L}_G$.
  end
```

CE+CH and CE+FCE) at 64 × 64. Here, CI, MI, FT, CE, CH and FCE stand for CIFAR10, MNIST, FASHION MNIST, CELEBAHQ, LSUN CHURCH and VERTICAL FLIP CELEBAHQ, respectively. "A+B" means "dataset A perturbed with 5% dataset B".

Comparison to DDGAN:

As shown in Tab. 1, our model consistently maintains strong performance even when the outlier percentage in training datasets increases. While the outlier ratio in the training dataset escalates from 3% to 10%, RDUOT's FID only increases by around 3.55 points (from 3.43 to 6.98). In contrast, DDGAN's FID increases by more than 10 points (from 4.76 to 14.77), and the synthesized outlier ratio of RDUOT rises from 0.2% to 3.8% compared to DDGAN's increase from 3.2% to 9.8%.

	Synthesized Outlier		FID	
Perturb ratio	DDGAN	RDUOT	DDGAN	RDUOT
3%	3.2%	0.2%	4.76	3.43
5%	4.1%	1.7%	8.81	4.37
7%	6.9%	2.3%	9.55	5.17
10%	9.8%	3.8%	14.77	6.98

Table 1: Synthesized Outlier Ratios and FID of DDGAN and RDUOT on CIFAR10 (perturbed by MNIST) with varying outlier ratios.

	RDUOT	DDGAN
CE+FT	7.89	10.68
CE+MT	9.29	12.95
CE+CH	7.86	9.83
CE+FCE	5.99	6.48

Table 2: FID of DDGAN and RDUOT on CE+FT, CE+CH, CE+MT and CE+FCE.

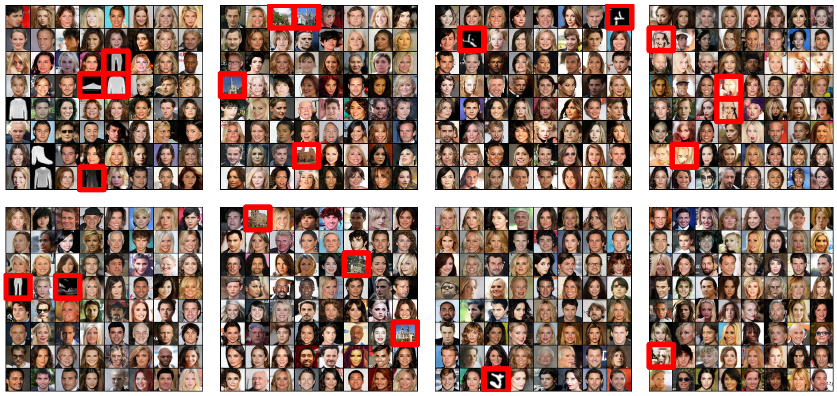


Fig. 1: From left to right is corresponding to CE+FT, CE+CH, CE+MT and CE+FCE dataset. Top: DDGAN, Bottom: RDUOT. The red boxes indicate the synthesized outliers among the clean synthesized samples.

When testing on higher dimensional datasets, RDUOT keeps dominating DDGAN as can be seen in Tab. 2 and Fig. 1. We observe that RDUOT performs well with both outlier datasets FT and MT which are grayscale and visually different from CE, with an FID gap of around 3 points when compared with the corresponding DDGAN model. Notably, even though the CH dataset comprises RGB images and bears great similarity to CE, RDUOT effectively learns to automatically eliminate outliers. For hard outlier dataset FCE, which has a great similarity with CE, RDUOT successfully removes the vertical flip face (refer to last column of Fig. 1) and we achieve a better FID score compared to DDGAN. This demonstrates RDUOT's capability to discriminate between two datasets in the same RGB domain, which has not previously been explored by other robust generative works [4, 6, 27].

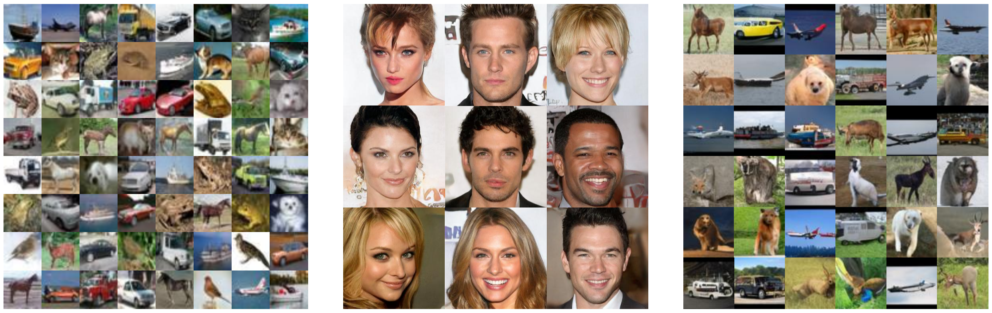
Comparison to other robust frameworks: As can be seen in Tab. 3, both UOTM [6] and RobustGAN [4] have much higher FID compared to RDUOT. RobustGAN is hard to converge and get very high FID even with two simple corrupted datasets. These results are even worse than DDGAN (Tab. 1). For UOTM, we first use KL as Ψ, but it cannot learn the data distribution and generate noisy images. We then use Softplus instead and got the FID reported in Tab. 3. However, UOTM still has a lower score compared to RDUOT. Specifically, the FID of UOTM on CE + FCE is higher than DDGAN's FID as shown in Tab. 1. These results prove the inferiority of the two existing models compared to RDUOT.

5.2 Performance in clean datasets

We assess the performance of RDUOT technique on three distinct clean datasets: CELEBA-HQ (256×256) [19], CIFAR-10 (32×32) [25], and STL-10 (64×64) [7] for image synthesis tasks. To assess the effectiveness of RDUOT, we utilize two

	CI+3%MT	CI+5%MT	CE+FT	CE+CH	CE+FCE
RDUOT	**3.43**	**4.37**	**7.89**	**7.86**	**5.99**
UOTM [6]	4.76	7.89	9.52	8.84	6.72
RobustGAN [4]	10.63	10.68	-	-	-

Table 3: Robustness comparison on CE+FT, CE+CH, CE+MT and CE+FCE. Note: RobustGAN uses the same architecture as UOTM and RDUOT for fair comparison.



(a) CIFAR-10 **(b)** CelebA-HQ **(c)** STL-10

Fig. 2: Qualitative results of RDUOT on 3 datasets STL-10, CIFAR-10, CelebA-HQ.

widely recognized metrics, namely FID [14] and Recall [26]. In Tab. 4 and Tab. 5, we can observe that RDUOT achieves significantly lower FID of **2.95** and **5.60** for CIFAR10 and CELEBA-HQ, in contrast to the baseline DDGAN, which records FID of 3.75 and 7.64 for CIFAR10 and CELEBA-HQ, respectively. Moreover, RDUOT achieves a better Recall of 0.58 compared to DDGAN's Recall of 0.57 for CIFAR10 and slightly outperforms DDGAN for CELEBA-HQ with a Recall of 0.38 compared to DDGAN's 0.36.

For STL-10 dataset, Tab. 6 illustrates a substantial improvement in FID for RDUOT compared to DDGAN. Specifically, RDUOT achieves a remarkable FID of **11.50**, roughly 10 points lower than DDGAN's FID of 21.79. Additionally, RDUOT achieves a higher Recall of 0.49, surpassing DDGAN's Recall of 0.40. Furthermore, RDUOT also outperforms all state-of-the-art methods in terms of FID and Recall.

In summary, our proposed RDUOT method outperforms the baseline DDGAN in **high-fidelity image generation** and maintains **good mode coverage**. In Tab. 7, we demonstrate that RDUOT **converges much faster** than DDGAN. By epoch 400, RDUOT achieves an FID of less than 20, while DDGAN's FID remains above 100. According to [29], in training process, stochastic diffusion process can go out of the support boundary, make itself diverge, and thus can generate highly unnatural samples. We hypothesize that the RDUOT's ability to remove outliers at each step (caused by the high variance of large diffusion

Model	FID↓	Recall↑	NFE↓
RDUOT	**2.95**	**0.58**	4
WaveDiff [36]	4.01	0.55	4
DDGAN [55]	3.75	0.57	4
DDPM [15]	3.21	0.57	1000
StyleGAN2 [22]	8.32	0.41	1
WGAN-GP [13]	39.40	-	1
RobustGAN [4]	21.57	-	1
RobustGAN*	11.40	-	1
OTM [39]	21.78	-	1
UOTM [6]	2.97	-	1
UOTM$^{\#}$	3.79	-	1

Table 4: Quantitative results on CIFAR-10. *: DDGAN architecture, $^{\#}$: trained on our machine

Model	FID↓	Recall↑
RDOUT	**5.60**	**0.38**
WaveDiff [36]	5.94	0.37
DDGAN [55]	7.64	0.36
Score SDE [48]	7.23	-
LFM [9]	5.26	-
NVAE [50]	29.7	-
VAEBM [54]	20.4	-
PGGAN [19]	8.03	-
VQ-GAN [11]	10.2	-
UOTM [6]	5.80	-

Table 5: Quantitative results on CELEBA-HQ.

Model	FID↓	Recall↑
Our	**11.50**	**0.49**
WaveDiff [36]	12.93	0.41
DDGAN [55]	21.79	0.40
StyleFormer [33]	15.17	-
TransGAN [18]	18.28	-
SNGAN [32]	40.1	-
StyleGAN2+ADA [20]	13.72	0.36
StyleGAN2+Aug [20]	12.97	0.39
Diffusion StyleGAN2 [53]	11.53	-

Table 6: Quantitative performance of RDUOT on STL-10. RDUOT surpasses DDGAN at both metric FID and Recall.

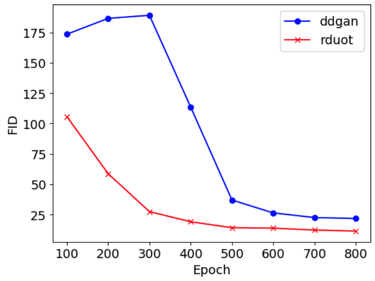


Table 7: The training convergence on STL-10 between DDGAN and RDUOT.

steps in DDGAN) leads to better performance. For a visual representation of our results, please refer to Fig. 2.

5.3 Ablation Study

Selection of Ψ:

Given that D_{Ψ_i} could be Csiszár-divergences, we can choose commonly used functions like KL and χ^2 for Ψ_1 and Ψ_2 in RDUOT. However, using KL as Ψ_i led to infinite loss during RDUOT training, even with meticulous hyperparameter tuning, likely due to the exponential convex conjugate form of KL (refer to Appendix 9). On clean CIFAR-10 dataset, using KL as Ψ, we obtain the best FID of 10.11 at epoch 1301 before the loss explodes to ∞. This phenomenon shows the instability of KL. For χ^2 as Ψ_i, the first row of Tab. 8 reveals that RDUOT with χ^2 achieve a FID score of 5.04, outperforming DDGAN's FID of 8.81 on CIFAR-10 with 5% outlier MNIST but still higher than softplus (4.37).

Ψ_1^*	Ψ_2^*	FID (clean) ↓	FID (5%) ↓
χ^2	χ^2	3.93	5.04
softplus	softplus	**2.95**	**4.37**

Table 8: FID for different choices of Ψ_1^* and Ψ_2^*.

Outlier ratio	0%	5%
Our	2.95	4.37
Our*	3.09	6.94
Our$^{\#}$	3.94	5.93

Table 9: Different proposed UOT losses. Our: Eq. (16), Our*: Eq. (13), Our$^{\#}$: Eq. (15)

Verifying Design of Framework:

In this section, we run experiments with other versions of our proposed model for verifying our insight in Sec. 4.1. The first version uses Eq. (13), and the second version uses Eq. (15). Their empirical results are shown in Tab. 9. Since noisy clean and outlier distributions at time t are close to each other, the proposed model using Eq. (13) fails to remove outliers (FID 6.94 compared to 4.37 of the main version). On the other hand, if using Eq. (15), the training process loses the information about x_{t-1} and hurts the sampling process, leading to worse performance as shown in Tab. 9.

6 Conclusion

In this paper, we introduce the first diffusion framework for robust-to-outliers image generation tasks. We present techniques to incorporate UOT into the DDGAN framework, leading to our proposed framework RDUOT. RDUOT has demonstrated the ability to either maintain or enhance performance across all three critical generative modeling criteria: mode coverage, high-fidelity generation, and fast sampling, all while ensuring rapid training convergence. Additionally, our paper showcases that RDUOT significantly outperforms DDGAN and other robust-to-outlier algorithms on corrupted training datasets with various settings, making it a promising approach for real-world corrupted datasets.

References

1. Akbari, A., Awais, M., Bashar, M., Kittler, J.: How does loss function affect generalization performance of deep learning? application to human age estimation. In: International Conference on Machine Learning. pp. 141–151. PMLR (2021)
2. Altschuler, J., Weed, J., Rigollet, P.: Near-linear time approximation algorithms for optimal transport via sinkhorn iteration. In: Advances in Neural Information Processing Systems. pp. 1964–1974 (2017)

3. Arjovsky, M., Chintala, S., Bottou, L.: Wasserstein generative adversarial networks. In: International conference on machine learning. pp. 214–223. PMLR (2017)
4. Balaji, Y., Chellappa, R., Feizi, S.: Robust optimal transport with applications in generative modeling and domain adaptation. In: NeurIPS (2020)
5. Chizat, L., Peyré, G., Schmitzer, B., Vialard, F.X.: Unbalanced optimal transport: Dynamic and kantorovich formulations. Journal of Functional Analysis **274**(11), 3090–3123 (2018)
6. Choi, J., Choi, J., Kang, M.: Generative modeling through the semi-dual formulation of unbalanced optimal transport. arXiv preprint arXiv:2305.14777 (2023)
7. Coates, A., Ng, A., Lee, H.: An analysis of single-layer networks in unsupervised feature learning. In: Gordon, G., Dunson, D., Dudík, M. (eds.) Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics. Proceedings of Machine Learning Research, vol. 15, pp. 215–223. PMLR, Fort Lauderdale, FL, USA (11–13 Apr 2011)
8. Cuturi, M.: Sinkhorn distances: Lightspeed computation of optimal transport. In: Advances in Neural Information Processing Systems. pp. 2292–2300 (2013)
9. Dao, Q., Phung, H., Nguyen, B., Tran, A.: Flow matching in latent space. arXiv preprint arXiv:2307.08698 (2023)
10. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. Advances in Neural Information Processing Systems **34**, 8780–8794 (2021)
11. Esser, P., Rombach, R., Ommer, B.: Taming transformers for high-resolution image synthesis (2020)
12. Gallouët, T., Ghezzi, R., Vialard, F.X.: Regularity theory and geometry of unbalanced optimal transport. arXiv preprint arXiv:2112.11056 (2021)
13. Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., Courville, A.C.: Improved training of wasserstein gans. Advances in neural information processing systems **30** (2017)
14. 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. Advances in neural information processing systems **30** (2017)
15. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: Advances in neural information processing systems (2020)
16. Huang, C.W., Lim, J.H., Courville, A.C.: A variational perspective on diffusion-based generative models and score matching. Advances in Neural Information Processing Systems **34**, 22863–22876 (2021)
17. Janati, H., Cuturi, M., Gramfort, A.: Spatio-temporal alignments: Optimal transport through space and time. arXiv preprint arXiv:1910.03860 (2019)
18. Jiang, Y., Chang, S., Wang, Z.: Transgan: Two pure transformers can make one strong gan, and that can scale up. Advances in Neural Information Processing Systems **34**, 14745–14758 (2021)
19. Karras, T., Aila, T., Laine, S., Lehtinen, J.: Progressive growing of GANs for improved quality, stability, and variation. In: International Conference on Learning Representations (2018)
20. Karras, T., Aittala, M., Hellsten, J., Laine, S., Lehtinen, J., Aila, T.: Training generative adversarial networks with limited data. In: Advances in neural information processing systems (2020)
21. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2019)

22. Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T.: Analyzing and improving the image quality of stylegan. In: Proceedings of the IEEE conference on computer vision and pattern recognition (2020)
23. Kingma, D., Salimans, T., Poole, B., Ho, J.: Variational diffusion models. Advances in neural information processing systems **34**, 21696–21707 (2021)
24. Kodali, N., Abernethy, J., Hays, J., Kira, Z.: On convergence and stability of gans. arXiv preprint arXiv:1705.07215 (2017)
25. Krizhevsky, A.: Learning multiple layers of features from tiny images. University of Toronto (05 2012)
26. Kynkäänniemi, T., Karras, T., Laine, S., Lehtinen, J., Aila, T.: Improved precision and recall metric for assessing generative models. Advances in Neural Information Processing Systems **32** (2019)
27. Le, K., Nguyen, H., Nguyen, Q., Ho, N., Pham, T., Bui, H.: On robust optimal transport: Computational complexity and barycenter computation (2021)
28. Le, T., Phung, H., Nguyen, T., Dao, Q., Tran, N., Tran, A.: Anti-dreambooth: Protecting users from personalized text-to-image synthesis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2023)
29. Lou, A., Ermon, S.: Reflected diffusion models. arXiv preprint arXiv:2304.04740 (2023)
30. Lu, C., Zhou, Y., Bao, F., Chen, J., Li, C., Zhu, J.: Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. arXiv preprint arXiv:2206.00927 (2022)
31. Meng, C., He, Y., Song, Y., Song, J., Wu, J., Zhu, J.Y., Ermon, S.: Sdedit: Guided image synthesis and editing with stochastic differential equations. arXiv preprint arXiv:2108.01073 (2021)
32. Miyato, T., Kataoka, T., Koyama, M., Yoshida, Y.: Spectral normalization for generative adversarial networks. arXiv preprint arXiv:1802.05957 (2018)
33. Park, J., Kim, Y.: Styleformer: Transformer based generative adversarial networks with style vector (2021)
34. Peyré, G., Cuturi, M.: Computational optimal transport. Foundations and Trends® in Machine Learning **11**(5-6), 355–607 (2019)
35. Pham, K., Le, K., Ho, N., Pham, T., Bui, H.: On unbalanced optimal transport: An analysis of sinkhorn algorithm (2020)
36. Phung, H., Dao, Q., Tran, A.: Wavelet diffusion models are fast and scalable image generators. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10199–10208 (June 2023)
37. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125 (2022)
38. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10684–10695 (2022)
39. Rout, L., Korotin, A., Burnaev, E.: Generative modeling with optimal transport maps. arXiv preprint arXiv:2110.02999 (2021)
40. Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation (2022)
41. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E., Ghasemipour, S.K.S., Ayan, B.K., Mahdavi, S.S., Lopes, R.G., et al.: Photorealistic text-to-image diffusion models with deep language understanding. arXiv preprint arXiv:2205.11487 (2022)

42. Salimans, T., Zhang, H., Radford, A., Metaxas, D.: Improving gans using optimal transport. arXiv preprint arXiv:1803.05573 (2018)
43. Sanjabi, M., Ba, J., Razaviyayn, M., Lee, J.D.: On the convergence and robustness of training gans with regularized optimal transport. Advances in Neural Information Processing Systems **31** (2018)
44. Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics. In: International Conference on Machine Learning (2015)
45. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. In: International Conference on Learning Representations (2021)
46. Song, Y., Durkan, C., Murray, I., Ermon, S.: Maximum likelihood training of score-based diffusion models. Advances in Neural Information Processing Systems **34**, 1415–1428 (2021)
47. Song, Y., Ermon, S.: Generative modeling by estimating gradients of the data distribution. In: Advances in neural information processing systems (2019)
48. Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-based generative modeling through stochastic differential equations. In: International Conference on Learning Representations (2021)
49. Vacher, A., Vialard, F.X.: Stability and upper bounds for statistical estimation of unbalanced transport potentials. arXiv preprint arXiv:2203.09143 (2022)
50. Vahdat, A., Kautz, J.: NVAE: A deep hierarchical variational autoencoder. In: Advances in neural information processing systems (2020)
51. Villani, C.: Optimal transport: Old and new (2008)
52. Wang, W., Bao, J., Zhou, W., Chen, D., Chen, D., Yuan, L., Li, H.: Semantic image synthesis via diffusion models. arXiv preprint arXiv:2207.00050 (2022)
53. Wang, Z., Zheng, H., He, P., Chen, W., Zhou, M.: Diffusion-gan: Training gans with diffusion. arXiv preprint arXiv:2206.02262 (2022)
54. Xiao, Z., Kreis, K., Kautz, J., Vahdat, A.: Vaebm: A symbiosis between variational autoencoders and energy-based models. In: International Conference on Learning Representations (2021)
55. Xiao, Z., Kreis, K., Vahdat, A.: Tackling the generative learning trilemma with denoising diffusion GANs. In: International Conference on Learning Representations (ICLR) (2022)
56. Yang, K.D., Uhler, C.: Scalable unbalanced optimal transport using generative adversarial networks. arXiv preprint arXiv:1810.11447 (2018)
57. Zhang, Q., Chen, Y.: Fast sampling of diffusion models with exponential integrator. arXiv preprint arXiv:2204.13902 (2022)