DPM-OT: A New Diffusion Probabilistic Model Based on Optimal Transport (Supplementary Material)

Zezeng Li[1,2], Shenghao Li[1], Zhanpeng Wang[3], Na Lei[1]*, Zhongxuan Luo[1], David Xianfeng Gu[4]
[1]School of Software, Dalian University of Technology, China
[2]Beijing Key Laboratory of Light-field Imaging and Digital Geometry, Capital Normal University, China
[3]School of Mathematical Sciences, University of the Chinese Academy of Sciences, China
[4]Computer Science and Applied Mathematics, State University of New York at Stony Brook, USA

In this document, we provide proof of theorems, comparisons with related works, additional implementation details, and experimental results.

1. Proof of Theorems

Theorem 1.1. *[4, 7] Given μ and ν on a compact convex domain $\Omega \subset \mathbb{R}^d$, there exists an OT plan for the cost $c(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{g}(\boldsymbol{x} - \boldsymbol{y})$, with $\boldsymbol{g}$ strictly convex. It is unique and of the form (id, $T_\#$) μ (id: identity map), provided that μ is absolutely continuous with respect to Lebesgue measure and $\partial\Omega$ is negligible. Moreover, there exists a Kantorovich's potential φ, and OT map T can be represented as follows:*

$$T(\boldsymbol{x}) = \boldsymbol{x} - (\nabla \boldsymbol{g})^{-1}[\nabla \varphi(\boldsymbol{x})].$$

Theorem 1.2. *Let $\tilde{\boldsymbol{x}}_t$ and $\boldsymbol{x}_t$ be the samples of step t obtained by* ***DPM-OT*** *and forward diffusion respectively, and $t \le M$, $\boldsymbol{\zeta}_M$ be the error at step M induced by the optimal trajectory, then there is a constant $C_t > 0$ satisfies the following inequality.*

$$\|\tilde{\boldsymbol{x}}_t - \boldsymbol{x}_t\| \le C_t \|\boldsymbol{\zeta}_M\| \tag{1}$$

Proof. Since the reverse diffusion function sequence $\{f_t\}$ is continuous, there is continuous function $c_t(\cdot)$ from $\mathbb{R}^d$ to $\mathbb{R}$ that makes the following formula hold

$$\begin{aligned} \tilde{\boldsymbol{x}}_t &= f_t \circ \cdots \circ f_{M-1}(\boldsymbol{x}_M + \boldsymbol{\zeta}_M) \\ &= f_t \circ \cdots \circ f_{M-1}(\boldsymbol{x}_M) + c_t(\boldsymbol{\zeta}_M)\boldsymbol{\zeta}_M \\ &= \boldsymbol{x}_t + c_t(\boldsymbol{\zeta}_M)\boldsymbol{\zeta}_M \end{aligned} \tag{2}$$

So we can get

$$\|\tilde{\boldsymbol{x}}_t - \boldsymbol{x}_t\| = \|c_t(\boldsymbol{\zeta}_M)\boldsymbol{\zeta}_M\| \le |c_t(\boldsymbol{\zeta}_M)| \, \|\boldsymbol{\zeta}_M\| \le C_t \|\boldsymbol{\zeta}_M\|$$

□

Theorem 1.3. *Let L_{dpm_ot} be the error between the data distribution generated by* ***DPM-OT*** *and the target data distribution which is defined in Eq. 3, L_{vlb} is the variational lower bound on negative log-likelihood between data distribution generated by vanilla DPM and the target data distribution which is defined in Eq. 4. We have $L_{dpm\_ot} \leqslant L_{vlb}$, i.e., L_{vlb} is the upper bound of L_{dpm_ot}.*

$$\begin{aligned} L_{dpm\_ot} =& L_0 + L_1 + ... + L_M + L_T \\ =& -\log \tilde{p}_\theta(\boldsymbol{x}_0|\boldsymbol{x}_1) + D_{KL}(q(\boldsymbol{x}_T|\boldsymbol{x}_0)), p(\boldsymbol{x}_T)) \\ &+ \sum_{t=1}^{M-1} D_{KL}(q(\boldsymbol{x}_t|\boldsymbol{x}_{t+1}, \boldsymbol{x}_0)||\tilde{p}_\theta(\boldsymbol{x}_t|\boldsymbol{x}_{t+1})) \\ &+ D_{KL}(q(\boldsymbol{x}_M|\boldsymbol{x}_T, \boldsymbol{x}_0)||\tilde{p}_\theta(\boldsymbol{x}_M|\boldsymbol{x}_T)), \end{aligned} \tag{3}$$

$$\begin{aligned} L_{vlb} =& -\log p_\theta(\boldsymbol{x}_0|\boldsymbol{x}_1) + D_{KL}(q(\boldsymbol{x}_T|\boldsymbol{x}_0))||p(\boldsymbol{x}_T)) \\ &+ \sum_{t>1} D_{KL}(q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0)||p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)) \end{aligned} \tag{4}$$

Proof. It may be assumed that$\boldsymbol{X}_t = \boldsymbol{x}_t - \mu_{t+1}(\boldsymbol{x}_{t+1})$,

$$\begin{aligned} p_\theta(\boldsymbol{x}_t|\boldsymbol{x}_{t+1}) &= \mathcal{N}(\boldsymbol{x}_t|\mu_{t+1}(\boldsymbol{x}_{t+1}), \sigma_{t+1}^2 I) \\ &= \frac{1}{(2\pi)^{d/2}\sigma_{t+1}^d} \exp\left(-\frac{1}{2\sigma_{t+1}^2}\boldsymbol{X}_t^T\boldsymbol{X}_t\right) \end{aligned} \tag{5}$$

$$\begin{aligned} & p_\theta(\tilde{\boldsymbol{x}}_t|\boldsymbol{x}_{t+1}) = \mathcal{N}(\boldsymbol{x}_t + \boldsymbol{\zeta}_t|\boldsymbol{\mu}_{t+1}(\boldsymbol{x}_{t+1}), \sigma_{t+1}^2 I) \\ &= \frac{1}{(2\pi)^{d/2}\sigma_{t+1}^d} \exp\left(-\frac{1}{2\sigma_{t+1}^2}(\boldsymbol{X}_t - \boldsymbol{\zeta}_t)^T(\boldsymbol{X}_t - \boldsymbol{\zeta}_t)\right) \end{aligned} \tag{6}$$

By definitions of $p_\theta(\boldsymbol{x}_t|\boldsymbol{x}_{t+1})$ and $p_\theta(\tilde{\boldsymbol{x}}_t|\boldsymbol{x}_{t+1})$, we know

*Corresponding author: *Na Lei (nalei@dlut.edu.cn)* the following equation

$$\begin{aligned}
&\left|\log\frac{p_\theta(\boldsymbol{x}_t|\boldsymbol{x}_{t+1})}{p_\theta(\tilde{\boldsymbol{x}}_t|\boldsymbol{x}_{t+1})}\right| \\
=&\frac{1}{2\sigma_{t+1}^2}\left|2\boldsymbol{x}_t\boldsymbol{\zeta}_t+\boldsymbol{\zeta}_t^T\boldsymbol{\zeta}_t-2\boldsymbol{\mu}_{t+1}(\boldsymbol{x}_{t+1})^T\boldsymbol{\zeta}_t\right| \\
\le&\frac{1}{2\sigma_{t+1}^2}\left(2||\boldsymbol{x}_t||\cdot||\boldsymbol{\zeta}_t||+||\boldsymbol{\zeta}_t||^2+2||\boldsymbol{\mu}_{t+1}(\boldsymbol{x}_{t+1})||\cdot\parallel\boldsymbol{\zeta}_t\parallel\right) \\
=&\frac{1}{2\sigma_{t+1}^2}\left(2||\boldsymbol{x}_t||+||\boldsymbol{\zeta}_t||+2||\boldsymbol{\mu}_{t+1}(\boldsymbol{x}_{t+1})||\right)||\boldsymbol{\zeta}_t||
\end{aligned} \tag{7}$$

Suppose $\boldsymbol{x}$ is bounded with $[a,b]^d$, then there is constant $A_t > 0$ makes

$$2||\boldsymbol{x}_t||+||\boldsymbol{\zeta}_t||+2||\boldsymbol{\mu}_{t+1}(\boldsymbol{x}_{t+1})|| \le A_t \tag{8}$$

Applying inequality 8 to equation 7, we get

$$\begin{aligned}
|\tilde{D}_{KL}-D_{KL}| =& \left|\int_{X_t} q(\boldsymbol{x}_t|\boldsymbol{x}_{t+1})\log\frac{p_\theta(\boldsymbol{x}_t|\boldsymbol{x}_{t+1})}{p_\theta(\tilde{\boldsymbol{x}}_t|\boldsymbol{x}_{t+1})}d\boldsymbol{x}_t\right| \\
\le&\frac{A_t}{2\sigma_{t+1}^2}\cdot||\boldsymbol{\zeta}_t||
\end{aligned} \tag{9}$$

where $\tilde{D}_{KL} = D_{KL}(q(\boldsymbol{x}_t|\boldsymbol{x}_{t+1})||p_\theta(\tilde{\boldsymbol{x}}_t|\boldsymbol{x}_{t+1}))$, $D_{KL} = D_{KL}(q(\boldsymbol{x}_t|\boldsymbol{x}_{t+1})||p_\theta(\boldsymbol{x}_t|\boldsymbol{x}_{t+1}))$. Therefore there is

$$\begin{aligned}
&|L_{dmot}-L_{vlb}^{0:M}| \\
\le&\sum_{t=0}^{M}|\tilde{D}_{KL}-D_{KL}| \\
\le&(M+1)\max_t\frac{A_t}{2\sigma_{t+1}^2}|C_t|\cdot||\boldsymbol{\zeta}_M||
\end{aligned} \tag{10}$$

That is

$$L_{dmot} \le L_{vlb}^{0:M}+(M+1)\max_t\frac{A_t}{2\sigma_{t+1}^2}|C_t|\cdot||\boldsymbol{\zeta}_M|| \tag{11}$$

Because of $\boldsymbol{\zeta}_M = O(N^{-\frac{1}{2}})$, we can make $\boldsymbol{\zeta}_M$ arbitrarily small by increasing the number of OT samples N. For a given **DPM**, there exists $\boldsymbol{\zeta}_M$ such that the following formula is true.

$$(M+1)\max_t\frac{A_t}{2\sigma_{t+1}^2}|C_t|\cdot||\boldsymbol{\zeta}_M|| \le L_{vlb}^{M+1:T} \tag{12}$$

So we have

$$L_{dmot} \le L_{vlb} \tag{13}$$

□

2. Comparisons with related works

This part mainly introduces three related works and their comparison to our method.

First, while the algorithm of this paper is inspired by Ref. [1], they differ in two main ways. 1) The algorithm for solving the OT map in our model does not need to obtain the latent codes using the autoencoder in advance; 2) As shown in Algorithm 1 of our paper, our algorithm is matrix-based, which is not trivial in engineering. In this paper, the $\hat{w}_i$ is calculated via the Monte Carlo (MC) method, which is estimated μ-volume of each cell W_i; $g(\cdot)$ is the OT map that transfers Gaussian noise to noisy image x_M.

Second, we analyze two acceleration models that employ truncated diffusion trajectories [8, 6] and compare them with our approach. Ref. [6] firstly uses an encoder to encode the image noise into a low dimension vector z, then uses a decoder to generate sample $x^{T'}$ from latent code z, finally, using the DDPM to denoise $x^{T'}$ to clean image in fewer steps; Ref. [8] utilizes an implicit generator to map the Gaussian noise to truncated noisy images and then applies the diffusion model for image generation; our work adopts an OT map to transfer the Gaussian noise to latent noisy images. Methods of Ref. [8, 6] require a possibly expensive training stage before they can be used for efficient sampling. They might require nontrivial effort to adapt the method to different models and the number of sampling steps. In contrast, our approach does not need to train additional network models and only requires the computation of the OT map. Secondly, under the guidance of OT theory, our method achieves higher-quality image generation with fewer reverse diffusion steps, especially in alleviating mode collapse.

3. Implementation Details

Additional details about hyperparameter settings of **Algorithm 1** and **Algorithm 2** are elucidated in this section. In the experiments, we instantiate the DPM model $\mathbf{s}_\theta$ of **DPM-OT** with pre-trained models of NCSNv2 [**?**]. In addition, for a fair comparison, we also adopted the same sampling schedule $\{(b_t,\sigma_t)\}_{t=0}^T$ as NCSNv2.

In **Algorithm 1**, we use the Monte Carlo method to solve the SDOT map. We set the number of Monte Carlo samples $N = 10 \times |\mathcal{I}|$, where $|\cdot|$ denotes the number of elements in the set. For the learning rate lr, we set the lr on the datasets CIFAR10 [3], CelebA [5] and FFHQ [2] to $lr = 0.1$, $lr = 20$, and $lr = 50$, respectively. For better convergence, we double the number of samples N and multiply the learning rate lr by 0.8 when the energy function $E(\boldsymbol{h})$ has not decreased for $s = 50$ steps. Moreover, we set threshold $\tau = 8 \times 10^{-4}$. When the energy function $E(\boldsymbol{h}) < \tau$ or the total number of iteration steps is greater than 10000, the optimization of $\boldsymbol{h}$ will be stopped.

In **Algorithm 2**, reverse diffusion steps M is a variable that satisfies $0 < M < T$. We have carried out five experiments on the datasets CIFAR10, CelebA, and FFHQ, where values of M are 5, 10, 20, 30 and 50 respectively. We find that with the increase of M, the image FID score will decrease, but this decline will tend to be flat with the increase of M, which is reflected in Table 1 of the paper.

4. Additional Results

In this section, we show three more quantitative and qualitative results of the proposed **DPM-OT** on FFHQ, Cifar10, and CelebA respectively.

For a fair comparison with baselines, we drew the NFE-FID curves in Fig. 4 which demonstrate the efficacy of our method in FID scores under the same NFE conditions.

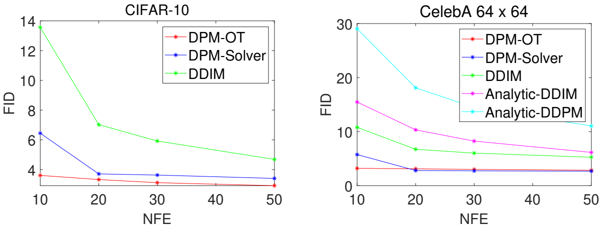

Figure 1. NFE-FID curves on CIFAR-10 and CelebA 64×64.

Furthermore, Fig. 2 and Fig. 3 show the results obtained by our model with 10 steps of inverse diffusion on the FFHQ 256×256 and Cifar10, respectively, and Fig. 4 shows the results obtained with only 5 steps of inverse diffusion on the CelebA dataset. As these results show, our model can obtain high-quality images after 5-10 reverse diffusion.

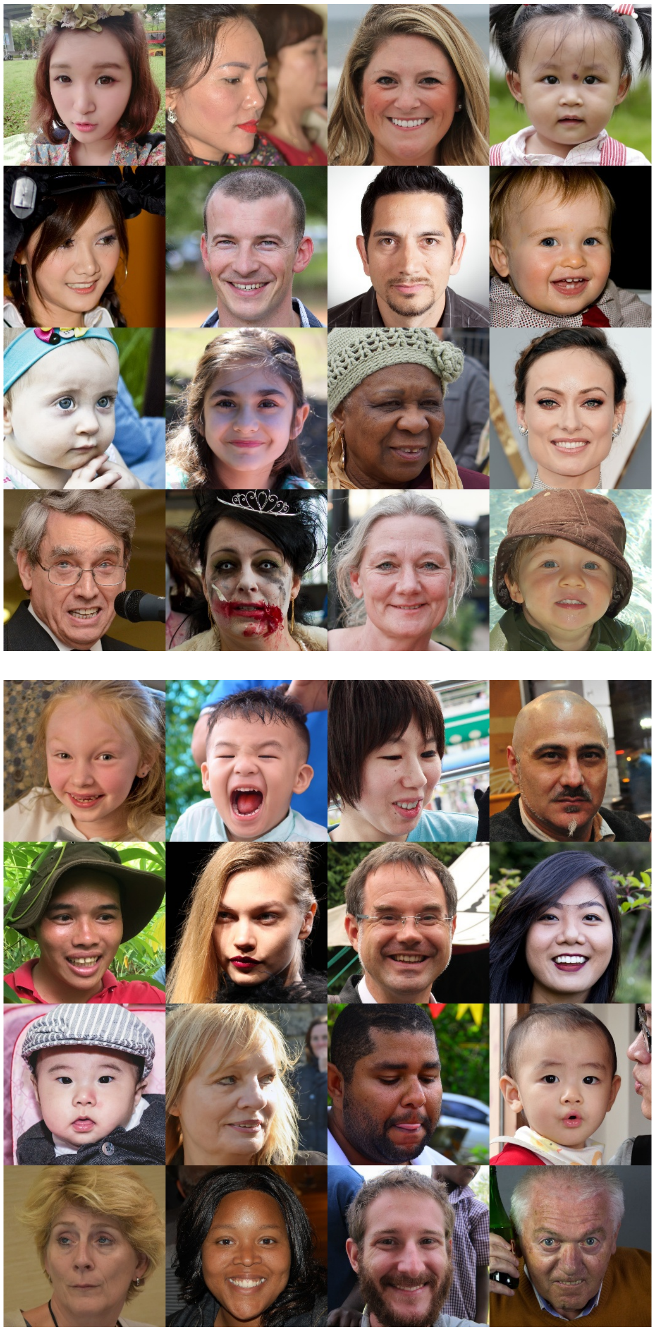

Figure 2. The visualization of our model on FFHQ (10 steps).

References

[1] Dongsheng An, Yang Guo, Na Lei, Zhongxuan Luo, Shing-Tung Yau, and Xianfeng Gu. Ae-ot: a new generative model based on extended semi-discrete optimal transport. *ICLR 2020*, 2019. 2

[2] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 2

[3] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 2

[4] Na Lei, Dongsheng An, Yang Guo, Kehua Su, Shixia Liu, Zhongxuan Luo, Shing-Tung Yau, and Xianfeng Gu. A geometric understanding of deep learning. *Engineering*, 6(3):361–374, 2020. 1

[5] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3730–3738, 2015. 2

[6] Zhaoyang Lyu, Xudong Xu, Ceyuan Yang, Dahua Lin, and Bo Dai. Accelerating diffusion models via early stop of the diffusion process. *arXiv preprint arXiv:2205.12524*, 2022. 2

[7] Cédric Villani. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2008. 1

[8] Huangjie Zheng, Pengcheng He, Weizhu Chen, and Mingyuan Zhou. Truncated diffusion probabilistic models and diffusion-based adversarial auto-encoders. *arXiv preprint arXiv:2202.09671*, 2022. 2

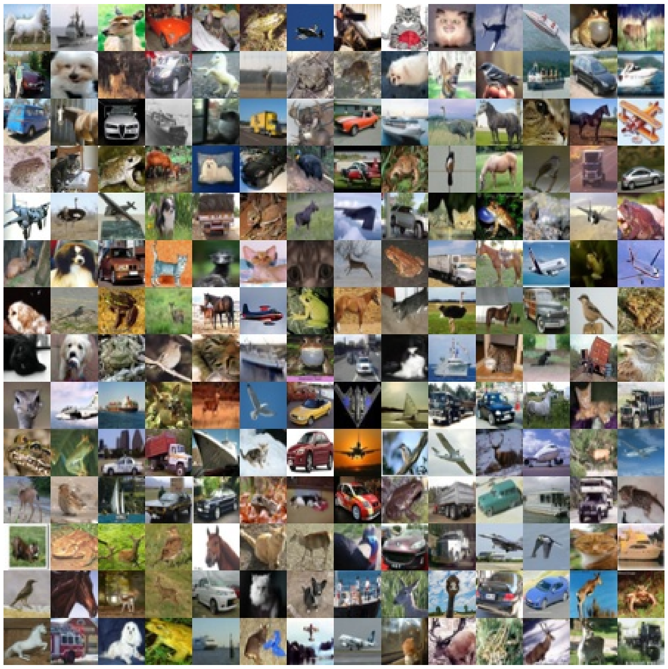

Figure 3. The visualization of our model on Cifar10 (10 steps).

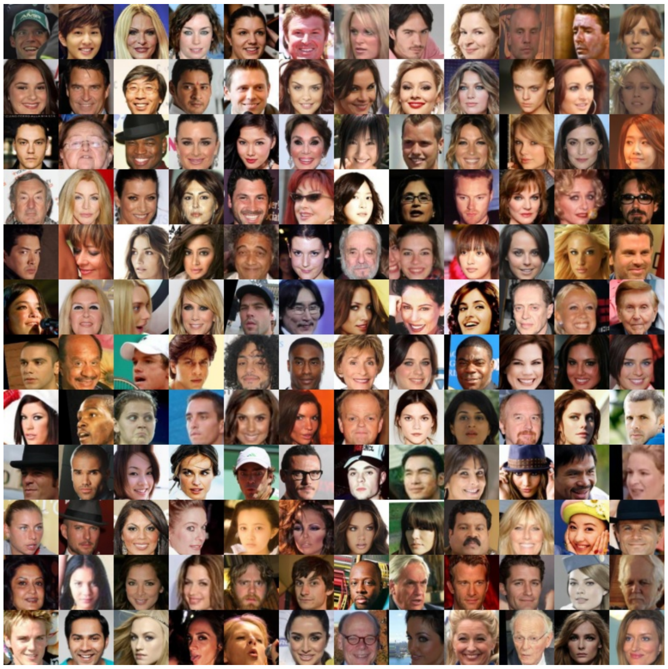

Figure 4. The visualization of our model on CelebA (5 steps).