Diffusion Models

Shiianov Vadim

March 17, 2022

Reparametrization trick

$$\xi \sim \mathcal{N}(\xi|\mu, \sigma^2) \Leftrightarrow \xi = \mu + \varepsilon\sigma,\ \varepsilon \sim \mathcal{N}(\varepsilon|0, 1)$$

Introduction

Paradigm	Quality	Diversity	Speed
VAE	✗	✓	✓
GAN	✓	✗	✓
Diffusion	✓	✓	✗

Figure: Table taken from `https://arankomatsuzaki.wordpress.com/2021/03/04/state-of-the-art-image-generative-models/`.

Deep Unsupervised Learning using Nonequilibrium Thermodynamics

Jascha Sohl-Dickstein JASCHA@STANFORD.EDU
Stanford University

Eric A. Weiss EAWEISS@BERKELEY.EDU
University of California, Berkeley

Niru Maheswaranathan NIRUM@STANFORD.EDU
Stanford University

Surya Ganguli SGANGULI@STANFORD.EDU
Stanford University

Abstract

A central problem in machine learning involves modeling complex data-sets using highly flexible families of probability distributions in which learning, sampling, inference, and evaluation are still analytically or computationally tractable. Here, we develop an approach that simultaneously achieves both flexibility and tractability. The essential idea, inspired by non-equilibrium statistical physics, is to systematically and slowly destroy structure in a data distribution through an iterative forward diffusion process. We then learn a reverse diffusion process that restores structure in data, yielding a highly flexible and tractable generative model of the data. This approach allows us to rapidly learn, sample from, these models are unable to aptly describe structure in rich datasets. On the other hand, models that are *flexible* can be molded to fit structure in arbitrary data. For example, we can define models in terms of any (non-negative) function $\phi(\mathbf{x})$ yielding the flexible distribution $p(\mathbf{x}) = \frac{\phi(\mathbf{x})}{Z}$, where Z is a normalization constant. However, computing this normalization constant is generally intractable. Evaluating, training, or drawing samples from such flexible models typically requires a very expensive Monte Carlo process.

A variety of analytic approximations exist which ameliorate, but do not remove, this tradeoff–for instance mean field theory and its expansions (T, 1982; Tanaka, 1998), variational Bayes (Jordan et al., 1999), contrastive divergence (Welling & Hinton, 2002; Hinton, 2002), minimum probability flow (Sohl-Dickstein et al., 2011b;a), minimum KL contraction (Lyu, 2011), proper scoring rules (Gneit-

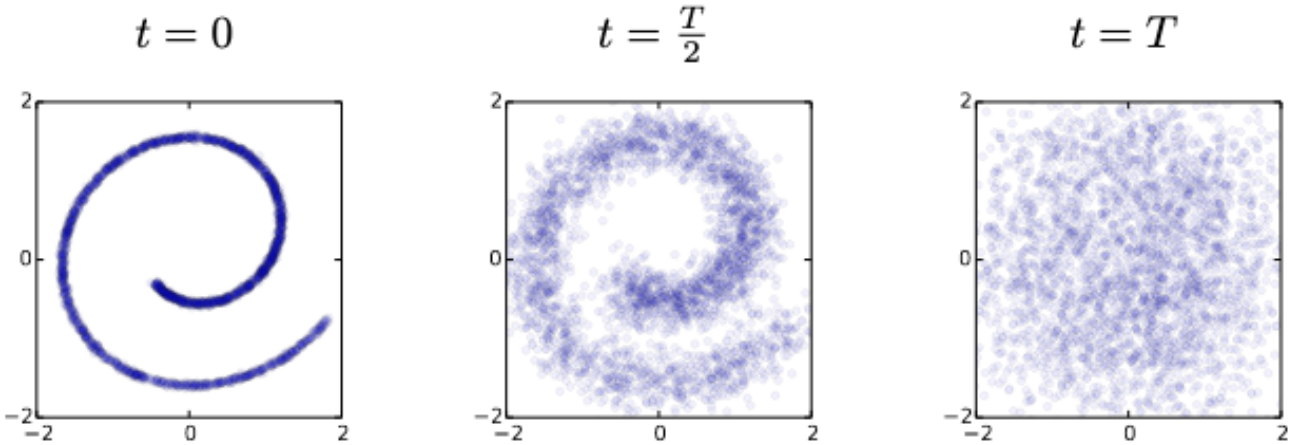
Diffusion Models: forward trajectory [4]

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

Diffusion Models: forward trajectory [4]

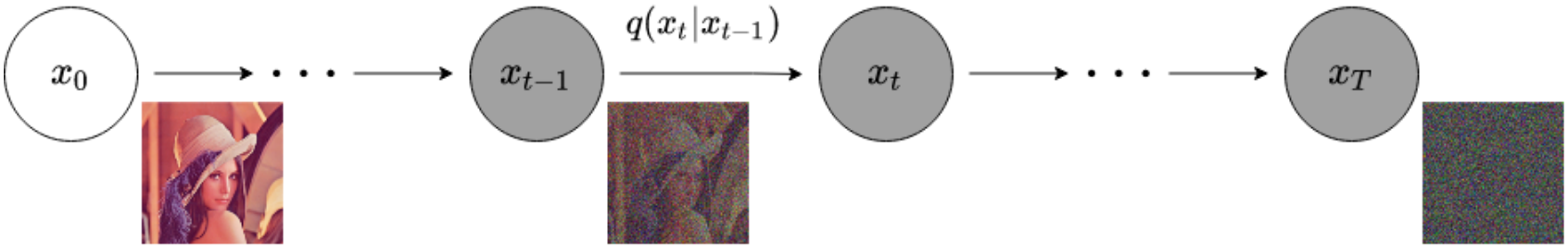
$$q(x_t|x_{t-1}) = \mathcal{N}(x_t|\sqrt{1-\beta_t}x_{t-1},\ \beta_t\mathbf{I})$$

$q\left(\mathbf{x}^{(0\cdots T)}\right)$



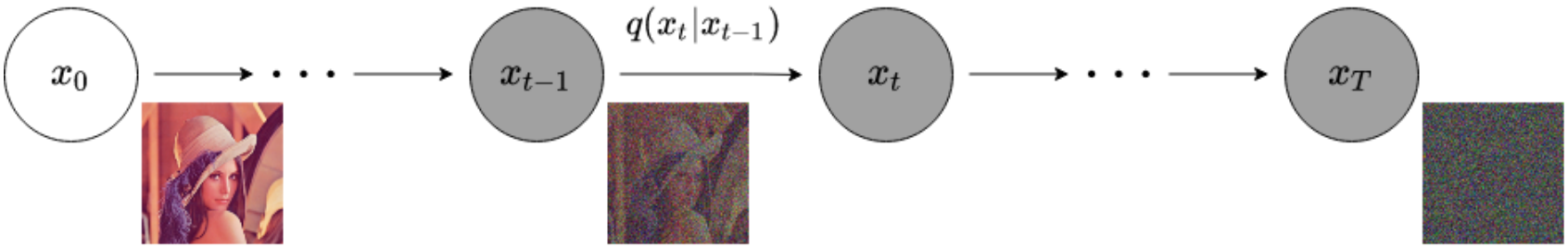
Diffusion Models: forward trajectory [4]

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



Diffusion Models: forward trajectory [4]

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

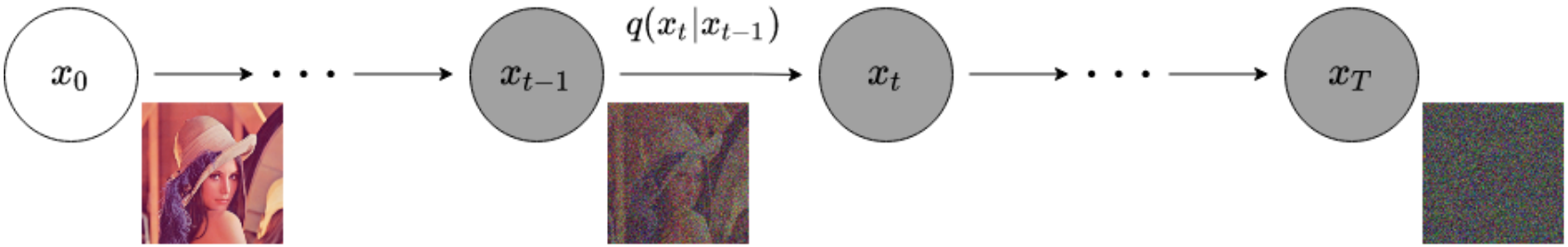


$$\alpha_t = 1 - \beta_t,\ \bar{\alpha}_t = \prod_{s=0}^{t} \alpha_s$$

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

Diffusion Models: forward trajectory [4]

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



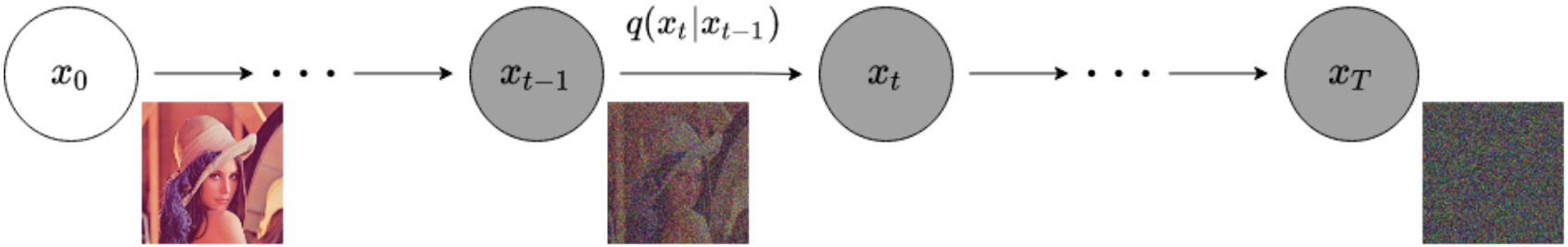
$$\alpha_t = 1-\beta_t,\ \bar{\alpha}_t = \prod_{s=0}^{t}\alpha_s$$

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

$$x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\varepsilon,\ \varepsilon \sim \mathcal{N}(\varepsilon|0,\ \mathbf{I})$$

Diffusion Models: forward trajectory [4]

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



$$q(x_{t-1}|x_t, x_0) = \mathcal{N}(x_{t-1}|\tilde{\mu}(x_t, x_0),\ \tilde{\beta}_t\mathbf{I})$$

$$\tilde{\mu}(x_t, x_0) = \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}x_0 + \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}x_t$$

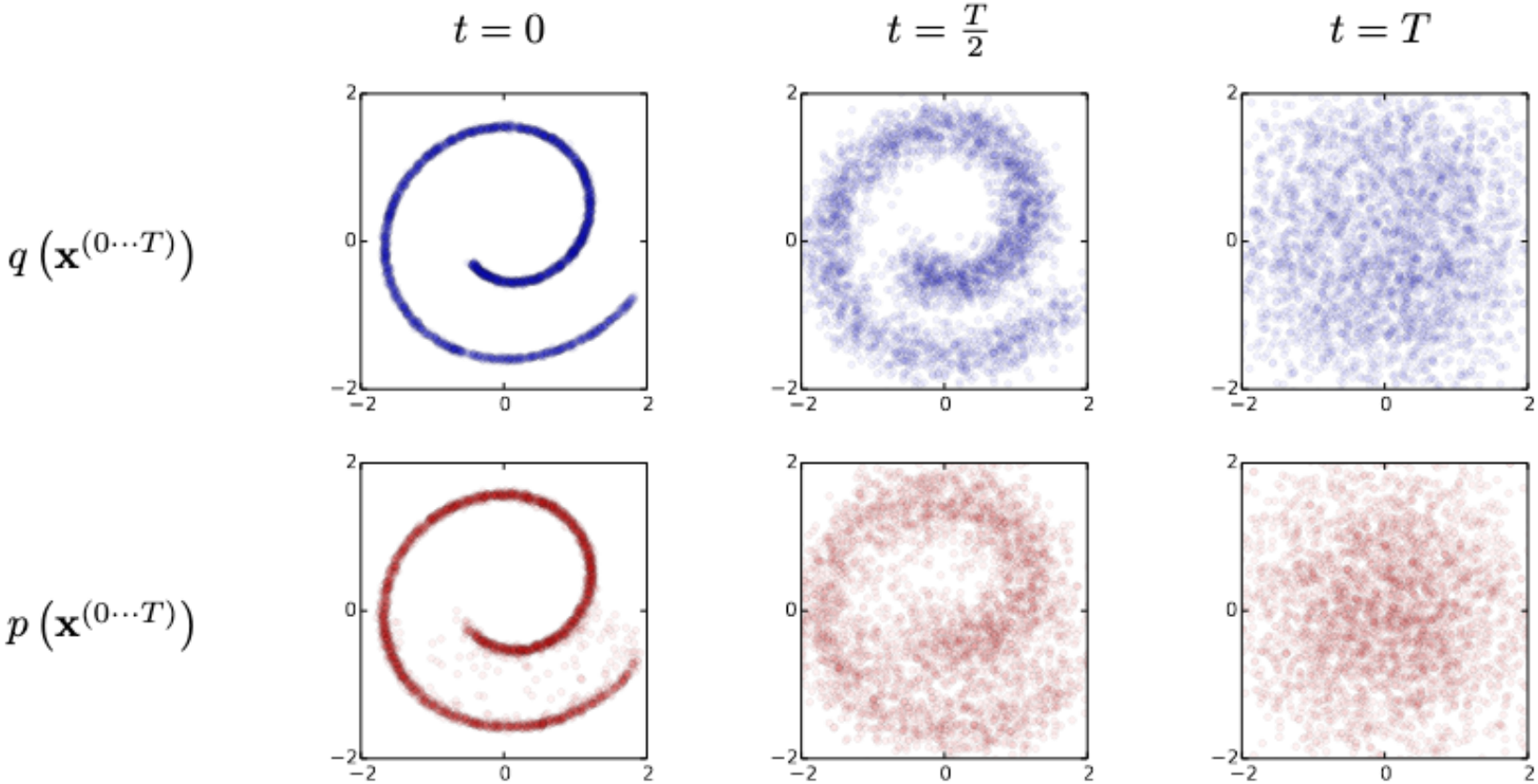
$$\tilde{\beta}_t = \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t$$

Diffusion Models: reverse trajectory [4]

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}|\mu_\theta(x_t, t),\ \Sigma_\theta(x_t, t))$$

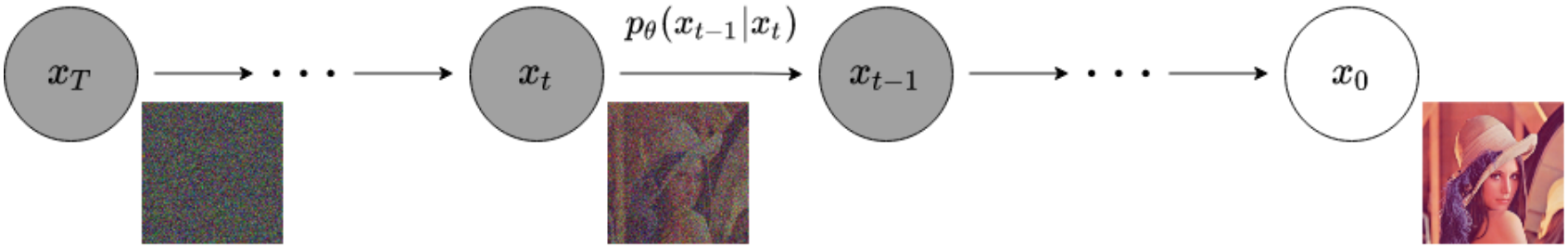
Diffusion Models: reverse trajectory [4]

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}|\mu_\theta(x_t, t),\ \Sigma_\theta(x_t, t))$$



Diffusion Models: reverse trajectory [4]

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}|\mu_\theta(x_t, t),\ \Sigma_\theta(x_t, t))$$



Diffusion Models: training objective [4]

$$\log p_\theta(x_0) = ???$$

Diffusion Models: training objective [4]

$$\log p_\theta(x_0) = ???$$

$$p_\theta(x_0, \ldots, x_T) = p(x_T) \prod_{t=1}^{T} p_\theta(x_{t-1}|x_t)$$

$$p_\theta(x_0) = \int p_\theta(x_0, \ldots, x_T) dx_1 \ldots dx_T$$

Diffusion Models: training objective [4]

$$p_\theta(x_0) = \int p_\theta(x_0, \ldots, x_T) dx_1 \ldots dx_T$$

Diffusion Models: training objective [4]

$$p_\theta(x_0) = \int p_\theta(x_0, \ldots, x_T) dx_1 \ldots dx_T =$$

$$= \int p_\theta(x_0, \ldots, x_T) \frac{q(x_1, \ldots, x_T | x_0)}{q(x_1, \ldots, x_T | x_0)} dx_1 \ldots dx_T =$$

$$= \int q(x_1, \ldots, x_T | x_0) \frac{p_\theta(x_0, \ldots, x_T)}{q(x_1, \ldots, x_T | x_0)} dx_1 \ldots dx_T$$

Diffusion Models: training objective [4]

$$p_\theta(x_0) = \int p_\theta(x_0, \ldots, x_T) dx_1 \ldots dx_T =$$

$$= \int p_\theta(x_0, \ldots, x_T) \frac{q(x_1, \ldots, x_T|x_0)}{q(x_1, \ldots, x_T|x_0)} dx_1 \ldots dx_T =$$

$$= \int q(x_1, \ldots, x_T|x_0) \frac{p_\theta(x_0, \ldots, x_T)}{q(x_1, \ldots, x_T|x_0)} dx_1 \ldots dx_T =$$

$$= \mathbb{E}_q \left[\frac{p_\theta(x_0, \ldots, x_T)}{q(x_1, \ldots, x_T|x_0)} \right]$$

Diffusion Models: training objective [4]

$$p_\theta(x_0) = \mathbb{E}_q \left[\frac{p_\theta(x_0, \ldots, x_T)}{q(x_1, \ldots, x_T | x_0)} \right]$$

Diffusion Models: training objective [4]

$$p_\theta(x_0) = \mathbb{E}_q \left[\frac{p_\theta(x_0, \ldots, x_T)}{q(x_1, \ldots, x_T | x_0)} \right]$$

$$\log p_\theta(x_0) = \log \mathbb{E}_{q(x_1, \ldots, x_T | x_0)} \left[\frac{p_\theta(x_0, \ldots, x_T)}{q(x_1, \ldots, x_T | x_0)} \right] \geq$$

$$\geq \mathbb{E}_{q(x_1, \ldots, x_T | x_0)} \left[\log \frac{p_\theta(x_0, \ldots, x_T)}{q(x_1, \ldots, x_T | x_0)} \right]$$

Diffusion Models: training objective [4]

$$\log p_\theta(x_0) \geq \mathbb{E}_{q(x_1,\ldots,x_T|x_0)} \left[\log \frac{p_\theta(x_0,\ldots,x_T)}{q(x_1,\ldots,x_T|x_0)} \right]$$

Diffusion Models: training objective [4]

$$\log p_\theta(x_0) \geq \mathbb{E}_{q(x_1,\ldots,x_T|x_0)} \left[\log \frac{p_\theta(x_0,\ldots,x_T)}{q(x_1,\ldots,x_T|x_0)} \right] =$$

$$= \mathbb{E}_q \left[\log \left(p(x_T) \prod_{t=1}^{T} \frac{p_\theta(x_{t-1}|x_t)}{q(x_t|x_{t-1})} \right) \right]$$

Diffusion Models: training objective [4]

$$\log p_\theta(x_0) \geq \mathbb{E}_{q(x_1,\ldots,x_T|x_0)}\left[\log \frac{p_\theta(x_0,\ldots,x_T)}{q(x_1,\ldots,x_T|x_0)}\right] =$$

$$= \mathbb{E}_q\left[\log\left(p(x_T)\prod_{t=1}^{T}\frac{p_\theta(x_{t-1}|x_t)}{q(x_t|x_{t-1})}\right)\right]$$

$$= \mathbb{E}_q\left[\log p(x_T) + \sum_{t=1}^{T}\log\frac{p_\theta(x_{t-1}|x_t)}{q(x_t|x_{t-1})}\right] =$$

$$= \mathbb{E}_q\left[\log p(x_T) + \sum_{t=2}^{T}\log\frac{p_\theta(x_{t-1}|x_t)}{q(x_t|x_{t-1})} + \log\frac{p_\theta(x_0|x_1)}{q(x_1|x_0)}\right]$$

Diffusion Models: training objective [4]

$$\log p_\theta(x_0) \geq \mathbb{E}_q \left[\log p(x_T) + \sum_{t=2}^{T} \log \frac{p_\theta(x_{t-1}|x_t)}{q(x_t|x_{t-1})} + \log \frac{p_\theta(x_0|x_1)}{q(x_1|x_0)} \right]$$

Diffusion Models: training objective [4]

$$\log p_\theta(x_0) \geq \mathbb{E}_q \left[\log p(x_T) + \sum_{t=2}^{T} \log \frac{p_\theta(x_{t-1}|x_t)}{q(x_t|x_{t-1})} + \log \frac{p_\theta(x_0|x_1)}{q(x_1|x_0)} \right]$$

$$q(x_t|x_{t-1}) = q(x_t|x_{t-1}, x_0)$$

$$q(x_t|x_{t-1}, x_0) = \frac{q(x_{t-1}|x_t, x_0)q(x_t|x_0)}{q(x_{t-1}|x_0)}$$

Diffusion Models: training objective [4]

$$\log p_\theta(x_0) \geq \mathbb{E}_q \left[\log p(x_T) + \sum_{t=2}^{T} \log \frac{p_\theta(x_{t-1}|x_t)}{q(x_t|x_{t-1})} + \log \frac{p_\theta(x_0|x_1)}{q(x_1|x_0)} \right]$$

$$q(x_t|x_{t-1}) = q(x_t|x_{t-1}, x_0)$$

$$q(x_t|x_{t-1}, x_0) = \frac{q(x_{t-1}|x_t, x_0) q(x_t|x_0)}{q(x_{t-1}|x_0)}$$

$$\sum_{t=2}^{T} \log \frac{p_\theta(x_{t-1}|x_t)}{q(x_t|x_{t-1})} = \sum_{t=2}^{T} \log \left(\frac{p_\theta(x_{t-1}|x_t)}{q(x_{t-1}|x_t, x_0)} \cdot \frac{q(x_{t-1}|x_0)}{q(x_t|x_0)} \right) =$$

$$= \sum_{t=2}^{T} \log \frac{p_\theta(x_{t-1}|x_t)}{q(x_{t-1}|x_t, x_0)} + \sum_{t=2}^{T} \log \frac{q(x_{t-1}|x_0)}{q(x_t|x_0)}$$

Diffusion Models: training objective [4]

$$\log p_\theta(x_0) \geq \mathbb{E}_q \left[\log p(x_T) + \sum_{t=2}^{T} \log \frac{p_\theta(x_{t-1}|x_t)}{q(x_t|x_{t-1})} + \log \frac{p_\theta(x_0|x_1)}{q(x_1|x_0)} \right]$$

$$\sum_{t=2}^{T} \log \frac{p_\theta(x_{t-1}|x_t)}{q(x_t|x_{t-1})} = \sum_{t=2}^{T} \log \frac{p_\theta(x_{t-1}|x_t)}{q(x_{t-1}|x_t, x_0)} + \sum_{t=2}^{T} \log \frac{q(x_{t-1}|x_0)}{q(x_t|x_0)}$$

Diffusion Models: training objective [4]

$$\log p_\theta(x_0) \geq \mathbb{E}_q \left[\log p(x_T) + \sum_{t=2}^{T} \log \frac{p_\theta(x_{t-1}|x_t)}{q(x_t|x_{t-1})} + \log \frac{p_\theta(x_0|x_1)}{q(x_1|x_0)} \right]$$

$$\sum_{t=2}^{T} \log \frac{p_\theta(x_{t-1}|x_t)}{q(x_t|x_{t-1})} = \sum_{t=2}^{T} \log \frac{p_\theta(x_{t-1}|x_t)}{q(x_{t-1}|x_t, x_0)} + \sum_{t=2}^{T} \log \frac{q(x_{t-1}|x_0)}{q(x_t|x_0)}$$

$$\sum_{t=2}^{T} \log \frac{q(x_{t-1}|x_0)}{q(x_t|x_0)} = \sum_{t=2}^{T} (\log q(x_{t-1}|x_0) - \log q(x_t|x_0))$$

Diffusion Models: training objective [4]

$$\log p_\theta(x_0) \geq \mathbb{E}_q \left[\log p(x_T) + \sum_{t=2}^{T} \log \frac{p_\theta(x_{t-1}|x_t)}{q(x_t|x_{t-1})} + \log \frac{p_\theta(x_0|x_1)}{q(x_1|x_0)} \right]$$

$$\sum_{t=2}^{T} \log \frac{p_\theta(x_{t-1}|x_t)}{q(x_t|x_{t-1})} = \sum_{t=2}^{T} \log \frac{p_\theta(x_{t-1}|x_t)}{q(x_{t-1}|x_t, x_0)} + \sum_{t=2}^{T} \log \frac{q(x_{t-1}|x_0)}{q(x_t|x_0)}$$

$$\sum_{t=2}^{T} \log \frac{q(x_{t-1}|x_0)}{q(x_t|x_0)} = \sum_{t=2}^{T} (\log q(x_{t-1}|x_0) - \log q(x_t|x_0)) =$$

$$= \log q(x_1|x_0) - \log q(x_T|x_0)$$

Diffusion Models: training objective [4]

$$\log p_\theta(x_0) \geq \mathbb{E}_q \left[\log p(x_T) + \sum_{t=2}^{T} \log \frac{p_\theta(x_{t-1}|x_t)}{q(x_t|x_{t-1})} + \log \frac{p_\theta(x_0|x_1)}{q(x_1|x_0)} \right]$$

$$\sum_{t=2}^{T} \log \frac{p_\theta(x_{t-1}|x_t)}{q(x_t|x_{t-1})} = \sum_{t=2}^{T} \log \frac{p_\theta(x_{t-1}|x_t)}{q(x_{t-1}|x_t, x_0)} + \sum_{t=2}^{T} \log \frac{q(x_{t-1}|x_0)}{q(x_t|x_0)}$$

$$\sum_{t=2}^{T} \log \frac{q(x_{t-1}|x_0)}{q(x_t|x_0)} = \log q(x_1|x_0) - \log q(x_T|x_0)$$

Diffusion Models: training objective [4]

$$\log p_\theta(x_0) \geq \mathbb{E}_q \left[\log p(x_T) + \sum_{t=2}^{T} \log \frac{p_\theta(x_{t-1}|x_t)}{q(x_t|x_{t-1})} + \log \frac{p_\theta(x_0|x_1)}{q(x_1|x_0)} \right]$$

$$\sum_{t=2}^{T} \log \frac{p_\theta(x_{t-1}|x_t)}{q(x_t|x_{t-1})} = \sum_{t=2}^{T} \log \frac{p_\theta(x_{t-1}|x_t)}{q(x_{t-1}|x_t, x_0)} + \sum_{t=2}^{T} \log \frac{q(x_{t-1}|x_0)}{q(x_t|x_0)} =$$

$$= \sum_{t=2}^{T} \log \frac{p_\theta(x_{t-1}|x_t)}{q(x_{t-1}|x_t, x_0)} + \log q(x_1|x_0) - \log q(x_T|x_0)$$

Diffusion Models: training objective [4]

$$\log p_\theta(x_0) \geq \mathbb{E}_q \left[\log p(x_T) + \sum_{t=2}^{T} \log \frac{p_\theta(x_{t-1}|x_t)}{q(x_t|x_{t-1})} + \log \frac{p_\theta(x_0|x_1)}{q(x_1|x_0)} \right] =$$

$$= \mathbb{E}_q \left[\log \frac{p(x_T)}{q(x_T|x_0)} + \sum_{t=2}^{T} \log \frac{p_\theta(x_{t-1}|x_t)}{q(x_{t-1}|x_t, x_0)} + \log p_\theta(x_0|x_1) \right]$$

Diffusion Models: training objective [4]

$$\log p_\theta(x_0) \geq \mathbb{E}_q \left[\log p(x_T) + \sum_{t=2}^{T} \log \frac{p_\theta(x_{t-1}|x_t)}{q(x_t|x_{t-1})} + \log \frac{p_\theta(x_0|x_1)}{q(x_1|x_0)} \right] =$$

$$= \mathbb{E}_q \left[\log \frac{p(x_T)}{q(x_T|x_0)} + \sum_{t=2}^{T} \log \frac{p_\theta(x_{t-1}|x_t)}{q(x_{t-1}|x_t, x_0)} + \log p_\theta(x_0|x_1) \right] =$$

$$= -D_{KL}(q(x_T|x_0) \parallel p(x_T)) - \sum_{t=2}^{T} D_{KL}(q(x_{t-1}|x_t, x_0) \parallel p_\theta(x_{t-1}|x_t)) +$$

$$+ \log p_\theta(x_0|x_1)$$

Diffusion Models: training objective [4]

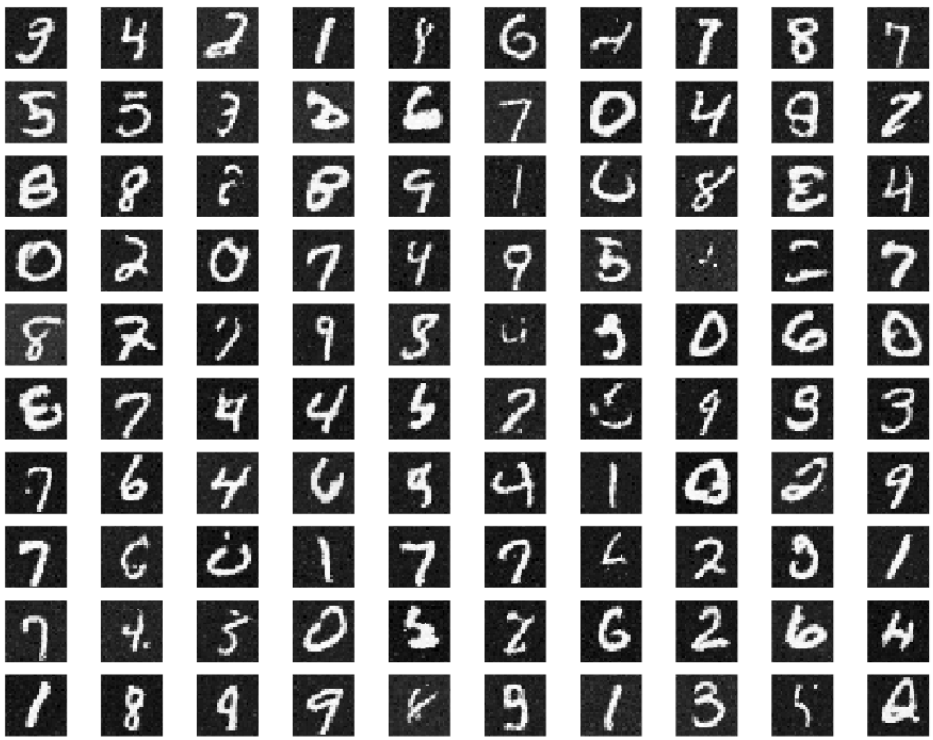
$$\mathbb{E}[-\log p_\theta(x_0)] \leq L_{\text{vlb}} = L_0 + \sum_{t=1}^{T} L_t$$

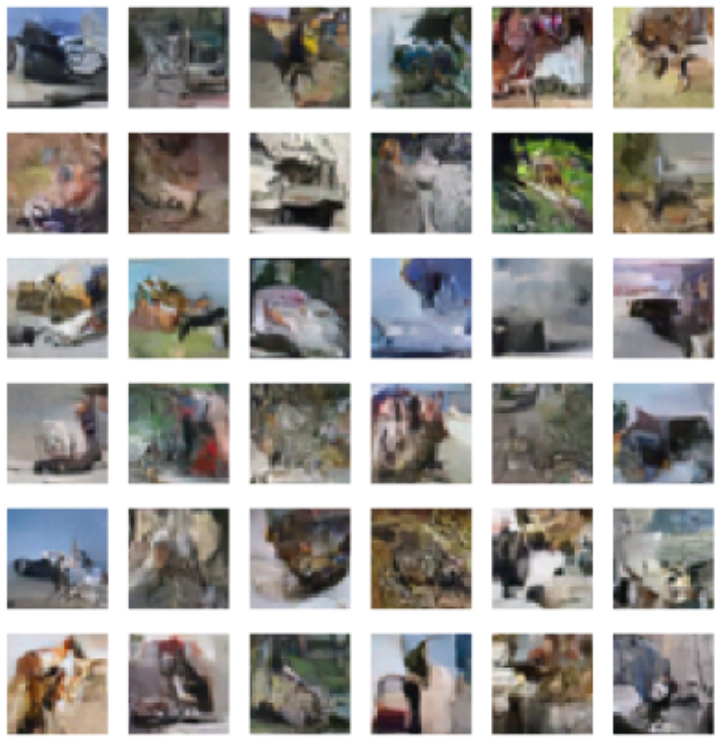
$$L_0 = -\log p_\theta(x_0|x_1)$$

$$L_{t-1} = D_{KL}(q(x_{t-1}|x_t, x_0) \parallel p_\theta(x_{t-1}|x_t))$$

$$L_T = D_{KL}(q(x_T|x_0) \parallel p(x_T))$$

Diffusion Models: results [4]





Denoising Diffusion Probabilistic Models [2]

$$x_t = \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1 - \bar{\alpha}_t} \varepsilon$$

Denoising Diffusion Probabilistic Models [2]

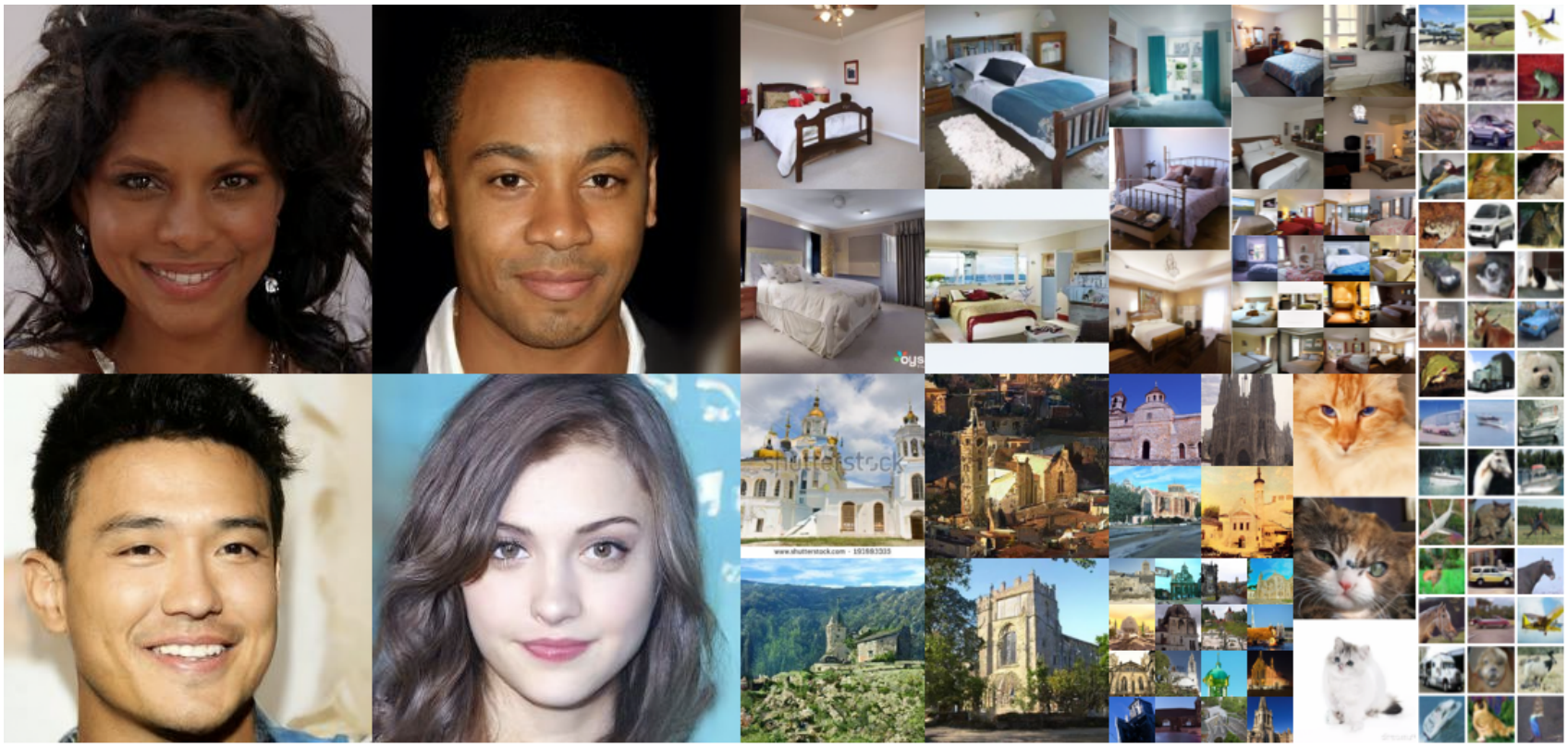
$$x_t = \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1 - \bar{\alpha}_t} \varepsilon$$

Denoising Diffusion Probabilistic Models [2]

$$x_t = \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1 - \bar{\alpha}_t} \varepsilon$$

$$L_{\text{simple}} = \mathbb{E}_{x_0 \sim q(x_0), t \sim [1,T], \varepsilon \sim \mathcal{N}(0,\mathbf{I})}[\|\varepsilon - \varepsilon_\theta(x_t, t)\|^2]$$

Denoising Diffusion Probabilistic Models [2]

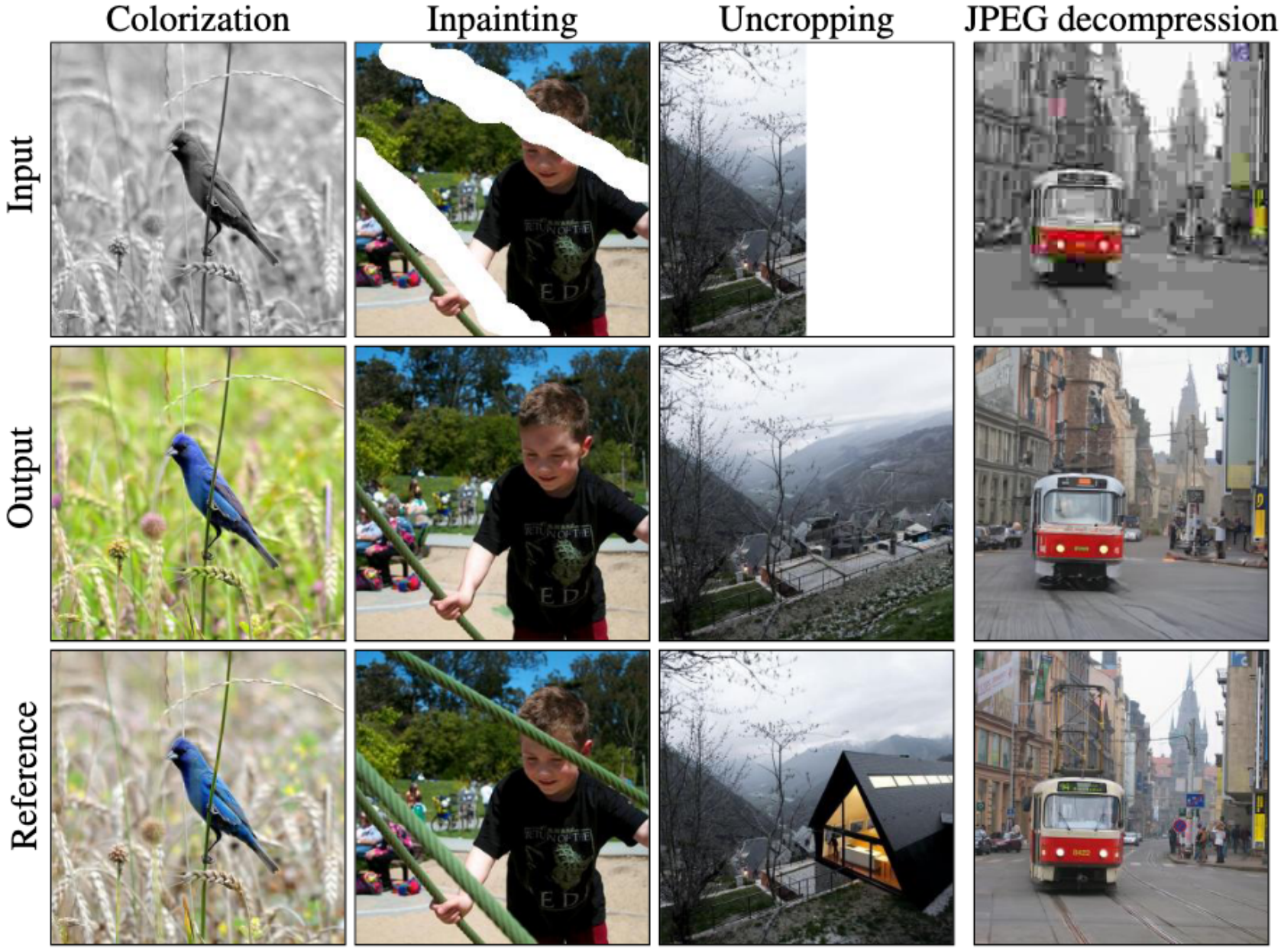


Diffusion Models Beat GANs on Image Synthesis [1]

- $\Sigma_\theta(x_t, t) = \exp(v(x_t, t) \log \beta_t + (1 - v(x_t, t)) \log \tilde{\beta}_t)$
- $L_{\text{simple}} + \lambda L_{\text{vlb}}$
- Adaptive Group Norm + Classifier Guidance

Diffusion Models Beat GANs on Image Synthesis [1]

Image-to-Image Diffusion Models [3]



My Contacts

Vadim Shiianov
E-Mail: vadimsh853@gmail.com
Telegram:
`https://t.me/binpord`
My Blog:
`https://binpord.github.io/`

References I

Prafulla Dhariwal and Alex Nichol.

Diffusion models beat gans on image synthesis.

Conference on Neural Information Processing Systems (NeurIPS), 2021, 2021.

Jonathan Ho, Ajay Jain, and Pieter Abbeel.

Denoising diffusion probabilistic models.

Conference on Neural Information Processing Systems (NeurIPS), 2020, 2020.

Chitwan Saharia, William Chan, Huiwen Chang, Chris A. Lee, Jonathan Ho, Tim Salimans, David J. Fleet, and Mohammad Norouzi.

Palette: Image-to-image diffusion models, 2021.

References II

Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli.

Deep unsupervised learning using nonequilibrium thermodynamics.

International Conference on Machine Learning (ICML), 2015, 2015.