ai.explained
INTUITIVE, PRACTICAL & MORE

An Introduction to Denoising Diffusion Probabilistic Model

2022/11/29

Applications

- Image generation

DDPM [Ho et.al, 2020]

Stable Diffusion [Rombach et.al, 2022]

Ho et.al, Denoising Diffusion Probabilistic Models, NeurIPS, 2020
Rombach et.al, High-Resolution Image Synthesis with Latent Diffusion Models, CVPR, 2022

Applications

- Image inpainting

Applications

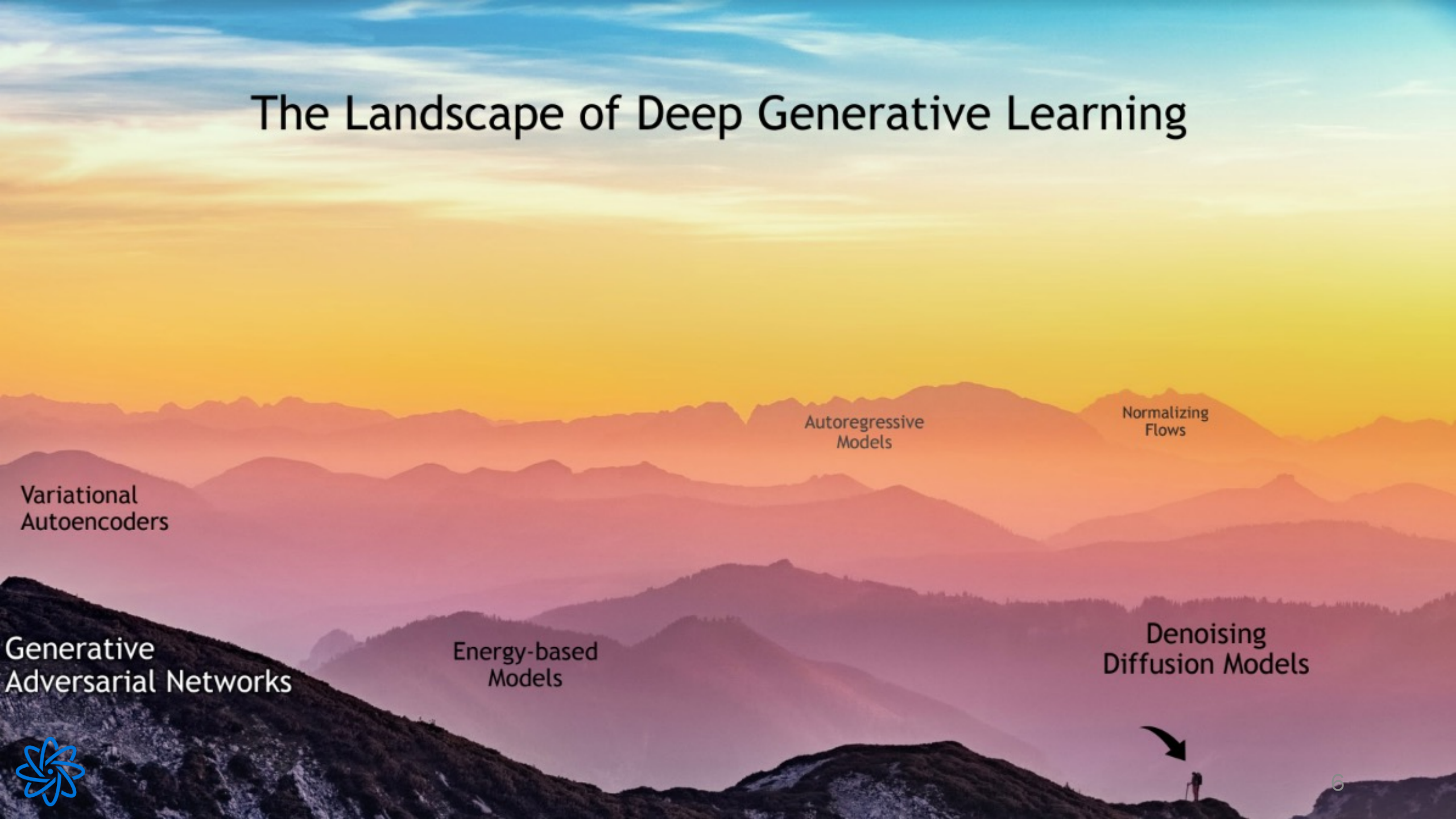
- Super resolution

Applications

- Text-to-image generation

“A high tech solarpunk utopia in the Amazon rainforest”

“a teddy bear on a skateboard in times square”


The Landscape of Deep Generative Learning
Autoregressive Models
Normalizing Flows
Variational Autoencoders
Generative Adversarial Networks
Energy-based Models
Denoising Diffusion Models

Denoising Diffusion Probabilistic Model

- Intuition

Destroy the structure

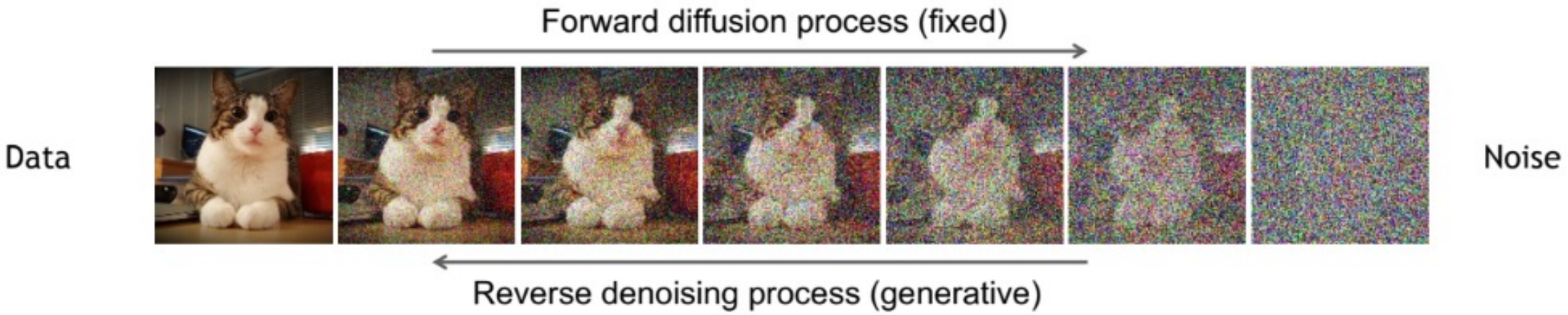
Data distribution → Uniform distribution

Data distribution ← Uniform distribution

Recover the structure

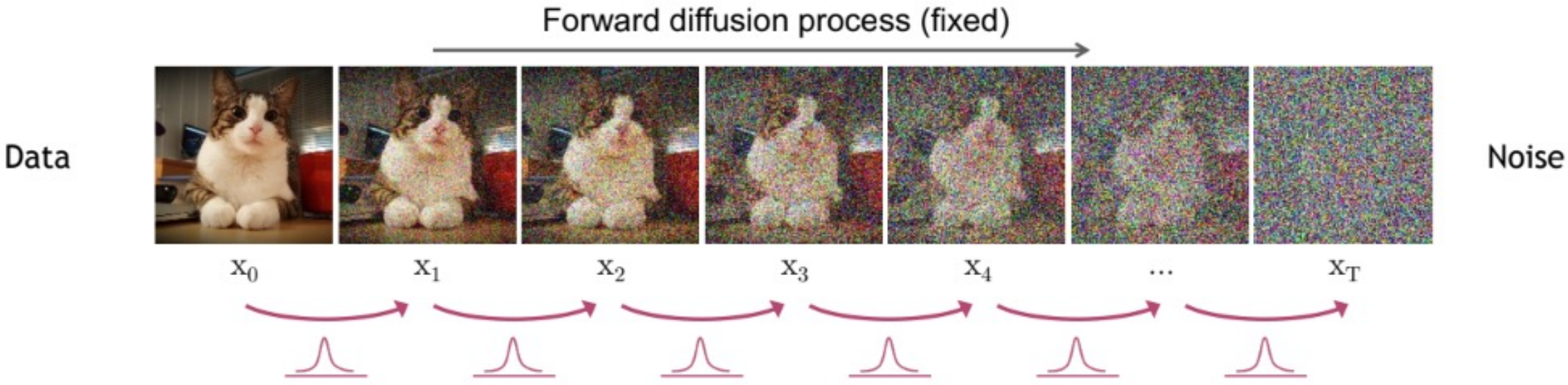
Denoising Diffusion Probabilistic Model

- Generate by denoising
  - Forward pass: gradually add noise
  - Reverse pass: gradually denoising



Forward pass

- Adding noise till totally destroyed



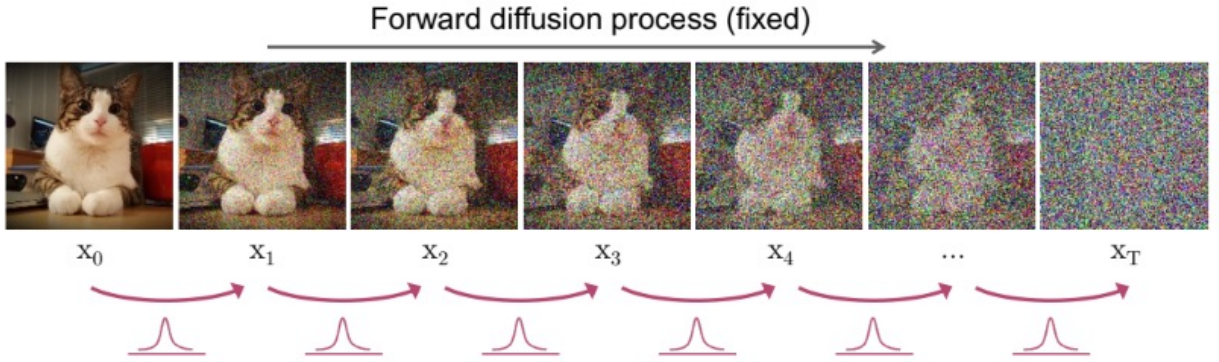
$$q(\mathbf{x}_t \mid \mathbf{x}_{t-1}) = \mathcal{N}\left(\mathbf{x_t}; \sqrt{1-\beta_t}\mathbf{x_{t-1}}, \beta_t\mathbf{I}\right) \Rightarrow q(\mathbf{x}_{1:T} \mid \mathbf{x}_0) = \prod_{t=1}^{T} q(\mathbf{x}_t \mid \mathbf{x}_{t-1})$$

Hyperparameter β_t is a small number, and increases as t grows. 0.95^100 = 0.005920529

Forward pass

- Adding noise till totally destroyed

$$q(\mathbf{x}_t \mid \mathbf{x}_{t-1}) = \mathcal{N}\left(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I}\right)$$



Let's write: $\alpha_t = 1 - \beta_t$, $\bar{\alpha}_t = \prod_{i=1}^{T} \alpha_i$,

$$\begin{aligned}
x_t &= \sqrt{a_t}x_{t-1} + \sqrt{1-\alpha_t}z_1 \quad \text{where} \quad z_1, z_2, \ldots \sim \mathcal{N}(0, \mathbf{I}) \\
&= \sqrt{a_t}\left(\sqrt{a_{t-1}}x_{t-2} + \sqrt{1-\alpha_{t-1}}z_2\right) + \sqrt{1-\alpha_t}z_1 \\
&= \sqrt{a_t a_{t-1}}x_{t-2} + \left(\sqrt{a_t(1-\alpha_{t-1})}z_2 + \sqrt{1-\alpha_t}z_1\right)
\end{aligned}$$

Recall that, $\mathcal{N}(0, \sigma_1^2\mathbf{I}) + \mathcal{N}(0, \sigma_2^2\mathbf{I}) \sim \mathcal{N}(0, (\sigma_1^2 + \sigma_2^2)\mathbf{I})$

$$\begin{aligned}
&\sqrt{a_t(1-\alpha_{t-1})}z_2 \sim \mathcal{N}(0, a_t(1-\alpha_{t-1})\mathbf{I}) \\
&\sqrt{1-\alpha_t}z_1 \sim \mathcal{N}(0, (1-\alpha_t)\mathbf{I}) \\
&\sqrt{a_t(1-\alpha_{t-1})}z_2 + \sqrt{1-\alpha_t}z_1 \sim \mathcal{N}(0, [\alpha_t(1-\alpha_{t-1}) + (1-\alpha_t)]\mathbf{I}) \\
&= \mathcal{N}(0, (1-\alpha_t\alpha_{t-1})\mathbf{I}).
\end{aligned}$$

$$\begin{aligned}
&= \sqrt{a_t a_{t-1}}x_{t-2} + \sqrt{1-\alpha_t\alpha_{t-1}}\bar{z}_2 \quad \text{where} \quad \bar{z}_2 \sim \mathcal{N}(0, \mathbf{I}) \\
&= \ldots \\
&= \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\bar{z}_t.
\end{aligned}$$

$$q(x_t \mid x_0) = \mathcal{N}\left(x_t; \sqrt{\bar{a}_t}x_0, (1-\bar{a}_t)\mathbf{I}\right)$$

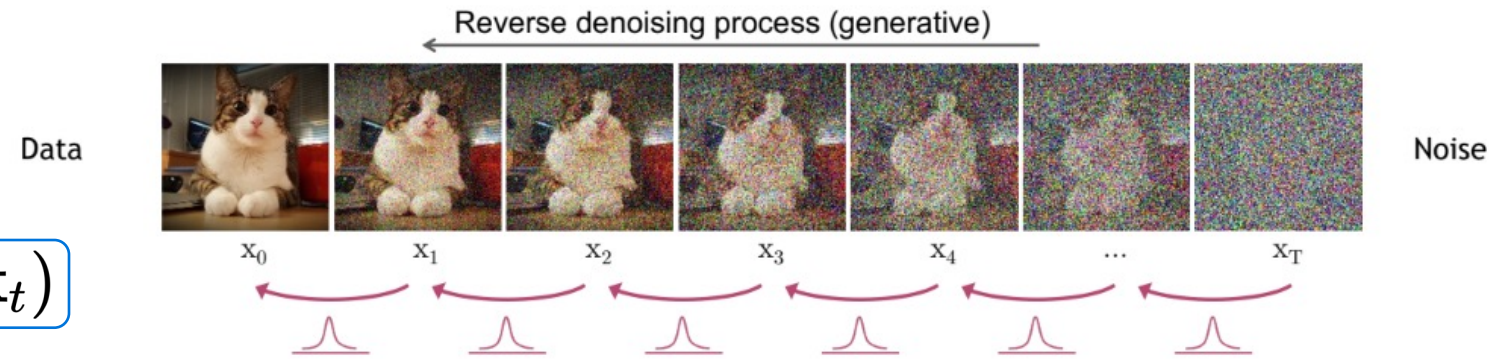
β_t is designed such that $q(\mathbf{x}_T \mid \mathbf{x}_0) \approx \mathcal{N}(\mathbf{x}_T; \mathbf{0}, \mathbf{I}))$

Reverse pass

- Denoising to generate data

Sample $\mathbf{x}_T \sim \mathcal{N}(\mathbf{x}_T; \mathbf{0}, \mathbf{I})$

Iteratively Sample $\mathbf{x}_{t-1} \sim q(\mathbf{x}_{t-1}|\mathbf{x}_t)$



True denoising distribution

$q(\mathbf{x}_{t-1}|\mathbf{x}_t) \propto q(\mathbf{x}_{t-1})q(x_t|\mathbf{x}_{t-1})$ is intractable, since huge mount of data need to be sampled. 😞

If β_t is small in forward pass, $q(\mathbf{x}_{t-1}|\mathbf{x}_t)$ would be a Normal distribution. [Sohl-Dickstein et.al, 2015]
So, why we not approximate $q(\mathbf{x}_{t-1}|\mathbf{x}_t)$? 😏

Sohl-Dickstein et.al, Deep unsupervised learning using nonequilibrium thermodynamics, ICML, 2015

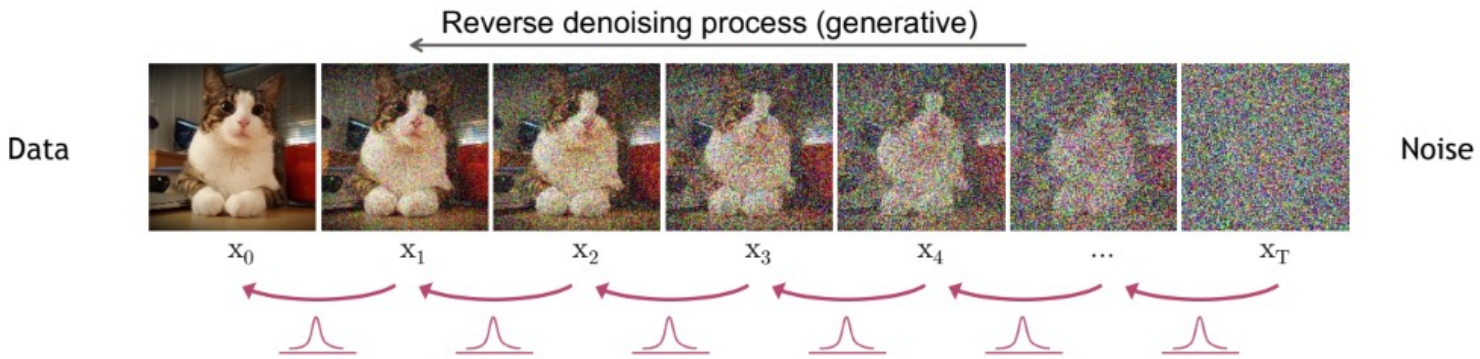
Reverse pass

- Denoising to generate data

$$p(\mathbf{x}_T) = \mathcal{N}(\mathbf{x}_T; \mathbf{0}, \mathbf{I}))$$
$$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}\left(\mathbf{x}_{t-1}; \mu_\theta(\mathbf{x}_t, t), \sigma_t^2\mathbf{I}\right)$$

U-Net

$$p_\theta(\mathbf{x}_{0:T}) = p(\mathbf{x}_T) \prod_{t=1}^{T} p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)$$



Learning Objective

- Maximize log-likelihood

$$\mathbb{E}_{q(\mathbf{x}_0)}[-\log p_\theta(\mathbf{x}_0)] \leq \mathbb{E}_{q(\mathbf{x}_0)}[-\log p_\theta(\mathbf{x}_0) + KL(q(\mathbf{x}_{1:T}|\mathbf{x}_0)||p_\theta(\mathbf{x}_{1:T}|\mathbf{x}_0))]$$

$$\leq \mathbb{E}_{q(\mathbf{x}_0)}\left[-\log p_\theta(\mathbf{x}_0) + \mathbb{E}_{q(\mathbf{x}_{1:T}|\mathbf{x}_0)} log \frac{q(\mathbf{x}_{1:T}|\mathbf{x}_0)}{p_\theta(\mathbf{x}_{1:T}|\mathbf{x}_0)}\right] \text{ (KL divergence definition)}$$

$$\leq \mathbb{E}_{q(\mathbf{x}_0)}\left[-\log p_\theta(\mathbf{x}_0) + \mathbb{E}_{q(\mathbf{x}_{1:T}|\mathbf{x}_0)}[log \frac{q(\mathbf{x}_{1:T}|\mathbf{x}_0)}{p_\theta(\mathbf{x}_{0:T})} + log p_\theta(\mathbf{x}_0)]\right]$$

$$\leq \mathbb{E}_{q(\mathbf{x}_0)}\left[\mathbb{E}_{q(\mathbf{x}_{1:T}|\mathbf{x}_0)} log \frac{q(\mathbf{x}_{1:T}|\mathbf{x}_0)}{p_\theta(\mathbf{x}_{0:T})}\right] =: L \text{ (variational upper bound)}$$

Pretty much the same as VAE!

Learning Objective

- Further derivation for the upper bound

$$
\begin{aligned}
L &:= \mathbb{E}_{q(\mathbf{x}_0)q(\mathbf{x}_{1:T}|\mathbf{x}_0)}\left[log\frac{q(\mathbf{x}_{1:T}|\mathbf{x}_0)}{p_\theta(\mathbf{x}_{0:T})}\right] \\
&= \mathbb{E}_{q(\mathbf{x}_0)q(\mathbf{x}_{1:T}|\mathbf{x}_0)}\left[log\frac{\prod_{t=1}^{T} q(\mathbf{x}_t|\mathbf{x}_{t-1})}{p_\theta(\mathbf{x}_T)\prod_{t=1}^{T} p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)}\right] \\
&= \mathbb{E}_{q(\mathbf{x}_0)q(\mathbf{x}_{1:T}|\mathbf{x}_0)}\left[-logp_\theta(\mathbf{x}_T) + \sum_{t=2}^{T} log[\frac{1}{p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)}q(\mathbf{x}_t|\mathbf{x}_{t-1})] + log\frac{q(\mathbf{x}_1|\mathbf{x}_0)}{p_\theta(\mathbf{x}_0|\mathbf{x}_1)}\right] \\
&= \mathbb{E}_{q(\mathbf{x}_0)q(\mathbf{x}_{1:T}|\mathbf{x}_0)}\left[-logp_\theta(\mathbf{x}_T) + \sum_{t=2}^{T} log[\frac{q(\mathbf{x}_{t-1}|\mathbf{x}_t,\mathbf{x}_0)}{p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)}\frac{q(\mathbf{x}_t|\mathbf{x}_0)}{q(\mathbf{x}_{t-1}|\mathbf{x}_0)}] + log\frac{q(\mathbf{x}_1|\mathbf{x}_0)}{p_\theta(\mathbf{x}_0|\mathbf{x}_1)}\right] \text{(Bayes' Rule)} \\
&= \mathbb{E}_{q(\mathbf{x}_0)q(\mathbf{x}_{1:T}|\mathbf{x}_0)}\left[-logp_\theta(\mathbf{x}_T) + \sum_{t=2}^{T} log\frac{q(\mathbf{x}_{t-1}|\mathbf{x}_t,\mathbf{x}_0)}{p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)} + log\frac{q(\mathbf{x}_T|\mathbf{x}_0)}{p_\theta(\mathbf{x}_0|\mathbf{x}_1)}\right]
\end{aligned}
$$

Learning Objective

- Further derivation for the upper bound (cont'd)

$$
\begin{aligned}
L &:= \mathbb{E}_{q(\mathbf{x}_0)q(\mathbf{x}_{1:T}|\mathbf{x}_0)}\left[log\frac{q(\mathbf{x}_{1:T}|\mathbf{x}_0)}{p_\theta(\mathbf{x}_{0:T})}\right] \\
&= \mathbb{E}_{q(\mathbf{x}_0)q(\mathbf{x}_{1:T}|\mathbf{x}_0)}\left[-logp_\theta(\mathbf{x}_T) + \sum_{t=2}^{T} log\frac{q(\mathbf{x}_{t-1}|\mathbf{x}_t,\mathbf{x}_0)}{p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)} + log\frac{q(\mathbf{x}_T|\mathbf{x}_0)}{p_\theta(\mathbf{x}_0|\mathbf{x}_1)}\right] \\
&= \mathbb{E}_{q(\mathbf{x}_0)q(\mathbf{x}_{1:T}|\mathbf{x}_0)}\left[log\frac{q(\mathbf{x}_T|\mathbf{x}_0)}{p_\theta(\mathbf{x}_T)} + \sum_{t=2}^{T} log\frac{q(\mathbf{x}_{t-1}|\mathbf{x}_t,\mathbf{x}_0)}{p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)} - logp_\theta(\mathbf{x}_0|\mathbf{x}_1)\right] \\
&= \mathbb{E}_{q(\mathbf{x}_0)q(\mathbf{x}_{1:T}|\mathbf{x}_0)}\Big[\underbrace{KL[q(\mathbf{x}_T|\mathbf{x}_0)||p_\theta(\mathbf{x}_T)]}_{\text{const}} + \sum_{t=2}^{T}\underbrace{KL[q(\mathbf{x}_{t-1}|\mathbf{x}_t,\mathbf{x}_0)||p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)]}_{\text{KL divergence}} \underbrace{- logp_\theta(\mathbf{x}_0|\mathbf{x}_1)}_{\text{Recon. Loss}}\Big] \\
&= \mathbb{E}_{q(\mathbf{x}_0)q(\mathbf{x}_{1:T}|\mathbf{x}_0)}[L_T + \sum_{t=2}^{T} L_t - logp_\theta(\mathbf{x}_0|\mathbf{x}_1)]
\end{aligned}
$$

Learning Objective

- Let's consider the $L_t = KL[q(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0) \| p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t)$

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

$$\propto \exp\left(-\frac{1}{2}\left(\frac{(x_t - \sqrt{\alpha_t} x_{t-1})^2}{\beta_t} + \frac{(x_{t-1} - \sqrt{\bar{\alpha}_{t-1}} x_0)^2}{1 - \bar{\alpha}_{t-1}} - \frac{(x_t - \sqrt{\bar{\alpha}_t} x_0)^2}{1 - \bar{\alpha}_t}\right)\right)$$

$$= \exp\left(-\frac{1}{2}\left(\underbrace{\left(\frac{\alpha_t}{\beta_t} + \frac{1}{1 - \bar{\alpha}_{t-1}}\right)}_{\text{mean of } x_1} x_{t-1}^2 - \underbrace{\left(\frac{2\sqrt{\alpha_t}}{\beta_t} x_t + \frac{2\sqrt{\bar{\alpha}_{t-1}}}{1 - \bar{\alpha}_{t-1}} x_0\right)}_{\text{variance of } x_{t-1}} x_{t-1} + \underbrace{C(x_t, x_0)}_{\text{const}}\right)\right)$$

In general, $\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) = \exp\left(-\frac{1}{2}\left(\frac{1}{\sigma^2}x^2 - \frac{2\mu}{\sigma^2}x + \frac{\mu^2}{\sigma^2}\right)\right)$

$$\frac{1}{\sigma^2} = \frac{1}{\tilde{\beta}_t} = \left(\frac{\alpha_t}{\beta_t} + \frac{1}{1-\bar{\alpha}_{t-1}}\right) \Rightarrow \quad \tilde{\beta}_t = \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t} \cdot \beta_t$$

$$\frac{2\mu}{\sigma^2} = \frac{2\tilde{\mu}_t(x_t, x_0)}{\tilde{\beta}_t} = \left(\frac{2\sqrt{\alpha_t}}{\beta_t} x_t + \frac{2\sqrt{\bar{\alpha}_{t-1}}}{1-\bar{\alpha}_{t-1}} x_0\right) \Rightarrow \tilde{\mu}_t(x_t, x_0) = \frac{\sqrt{a_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t} x_t + \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t} x_0.$$

$$q(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0) = \mathcal{N}\left(\mathbf{x}_{t-1}; \tilde{\mu}_t(\mathbf{x}_t, \mathbf{x}_0), \tilde{\beta}_t \mathbf{I}\right)$$

Learning Objective

- Let's consider the $L_t = KL[q(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0) \| p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t)$ (cont'd)

$$q(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0) = \mathcal{N}\left(\mathbf{x}_{t-1}; \tilde{\mu}_t(\mathbf{x}_t, \mathbf{x}_0), \tilde{\beta}_t \mathbf{I}\right)$$

$$p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t) = \mathcal{N}\left(\mathbf{x}_{t-1}; \mu_\theta(\mathbf{x}_t, t), \sigma_t^2 \mathbf{I}\right)$$

$$L_t = \mathrm{KL}(q(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0) \| p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t)) = \mathbb{E}_q \left[\frac{1}{2\sigma_t^2} \| \tilde{\mu}_t(\mathbf{x}_t, \mathbf{x}_0) - \mu_\theta(\mathbf{x}_t, t) \|^2 \right] + C$$

- Recall (slide #10) that $q(x_t \mid x_0) = \mathcal{N}\left(x_t; \sqrt{\bar{a}_t} x_0, (1 - \bar{a}_t)\mathbf{I}\right)$

$$\begin{aligned}
\tilde{\mu}_t(x_t, x_0) &= \frac{\sqrt{a_t}(1 - \bar{\alpha}_{t-1})}{1 - \bar{\alpha}_t} x_t + \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1 - \bar{\alpha}_t} x_0 \\
&= \frac{\sqrt{a_t}(1 - \bar{\alpha}_{t-1})}{1 - \bar{\alpha}_t} x_t + \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1 - \bar{\alpha}_t} \frac{x_t - \sqrt{1 - \bar{\alpha}_t}\,\epsilon}{\sqrt{\bar{\alpha}_t}} \\
&= \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{\beta_t}{\sqrt{1 - \bar{\alpha}_t}}\epsilon\right)
\end{aligned}$$

To predict noise instead

$$\mu_\theta(x_t, x_0) = \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{\beta_t}{\sqrt{1 - \bar{\alpha}_t}}\epsilon_\theta(x_t, t)\right)$$

DDPM [Ho et.al, 2020]

Learning Objective

- The final form

$$L_t = \mathrm{KL}(q(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0) \| p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t)) = \mathbb{E}_q \left[\frac{1}{2\sigma_t^2} \| \tilde{\mu}_t(\mathbf{x}_t, \mathbf{x}_0) - \mu_\theta(\mathbf{x}_t, t) \|^2 \right] + C$$

$$\tilde{\mu}_t(x_t, x_0) = \frac{1}{\sqrt{\alpha_t}} (x_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}} \epsilon) \qquad \mu_\theta(x_t, x_0) = \frac{1}{\sqrt{\alpha_t}} (x_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}} \epsilon_\theta(x_t, t))$$

$$\Rightarrow L_t = \mathbb{E}_{\mathbf{x}_0 \sim q(\mathbf{x}_0), \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})} \left[\frac{\beta_t^2}{2\sigma_t^2 \alpha_t (1-\bar{\alpha}_t)} ||\epsilon - \epsilon_\theta(\mathbf{x}_t, t)||^2 \right] + C$$

Loss weights

$$L_{simple} = \mathbb{E}_{\mathbf{x}_0 \sim q(\mathbf{x}_0), \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), t \sim \mathcal{U}(1, T)} \left[||\epsilon - \epsilon_\theta(\mathbf{x}_t, t)||^2 \right]$$ DDPM [Ho et.al, 2020]

Summary

- The training & sampling recipe

Algorithm 1 Training

1: **repeat**
2: $\mathbf{x}_0 \sim q(\mathbf{x}_0)$
3: $t \sim \text{Uniform}(\{1, \dots, T\})$
4: $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
5: Take gradient descent step on
$$\nabla_\theta \left\| \boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, t) \right\|^2$$
sampled x_t
6: **until** converged

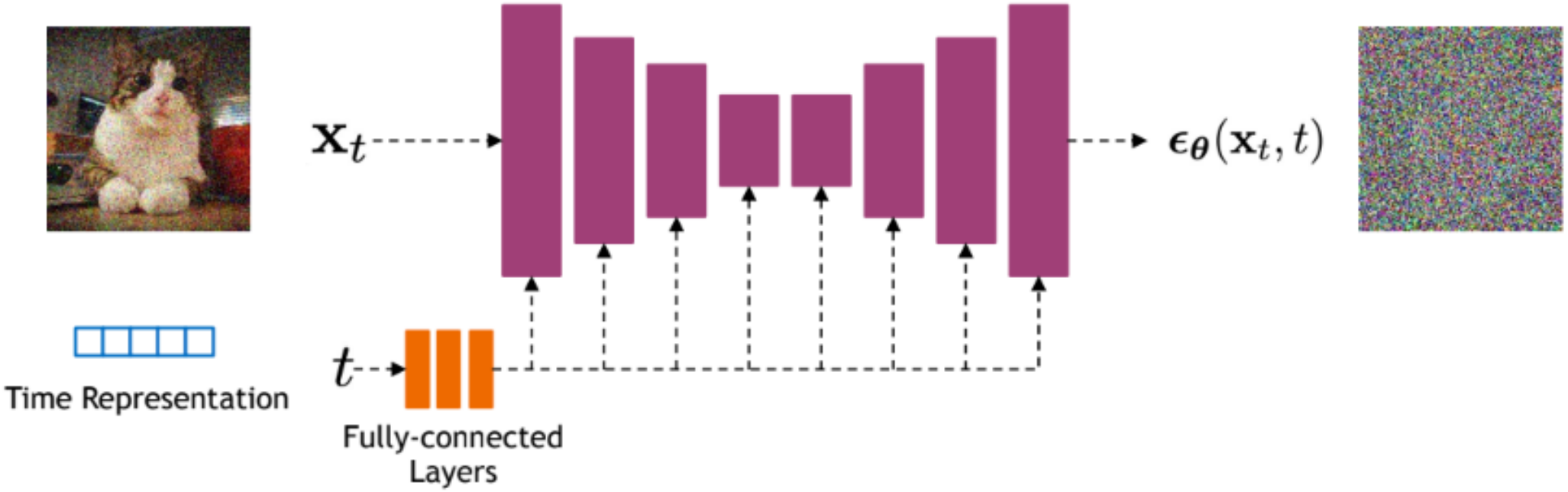
Algorithm 2 Sampling

1: $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
2: **for** $t = T, \dots, 1$ **do**
3: $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ if $t > 1$, else $\mathbf{z} = \mathbf{0}$
4: $\mathbf{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)\right) + \sigma_t \mathbf{z}$
Predicted mean
5: **end for**
6: **return** $\mathbf{x}_0$

$$L_{simple} = \mathbb{E}_{\mathbf{x}_0 \sim q(\mathbf{x}_0), \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), t \sim \mathcal{U}(1, T)} \left[||\epsilon - \epsilon_\theta(\mathbf{x}_t, t)||^2 \right]$$

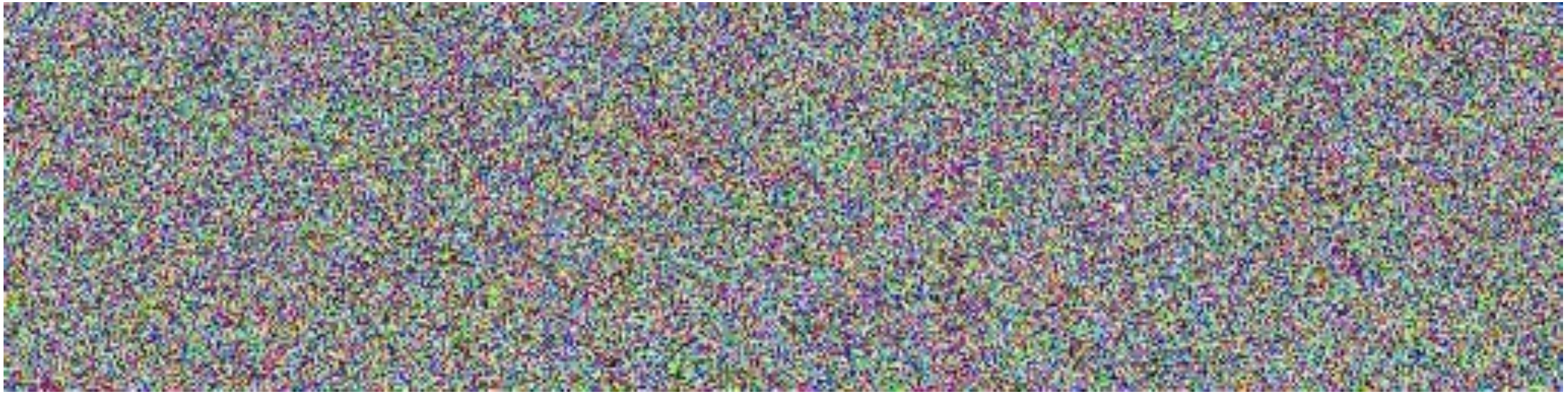
Implementation Details

- U-Net with time-embedding



Experiments Demo

- Trained on LSUN church



LSUN: https://www.yf.io/p/lsun

Thanks for watching!

For more videos, stay tuned. 🤓