Generative Modeling by Estimating Gradients of the Data Distribution

Yang Song, Stefano Ermon

Stanford University

NeurIPS 2019 (oral)

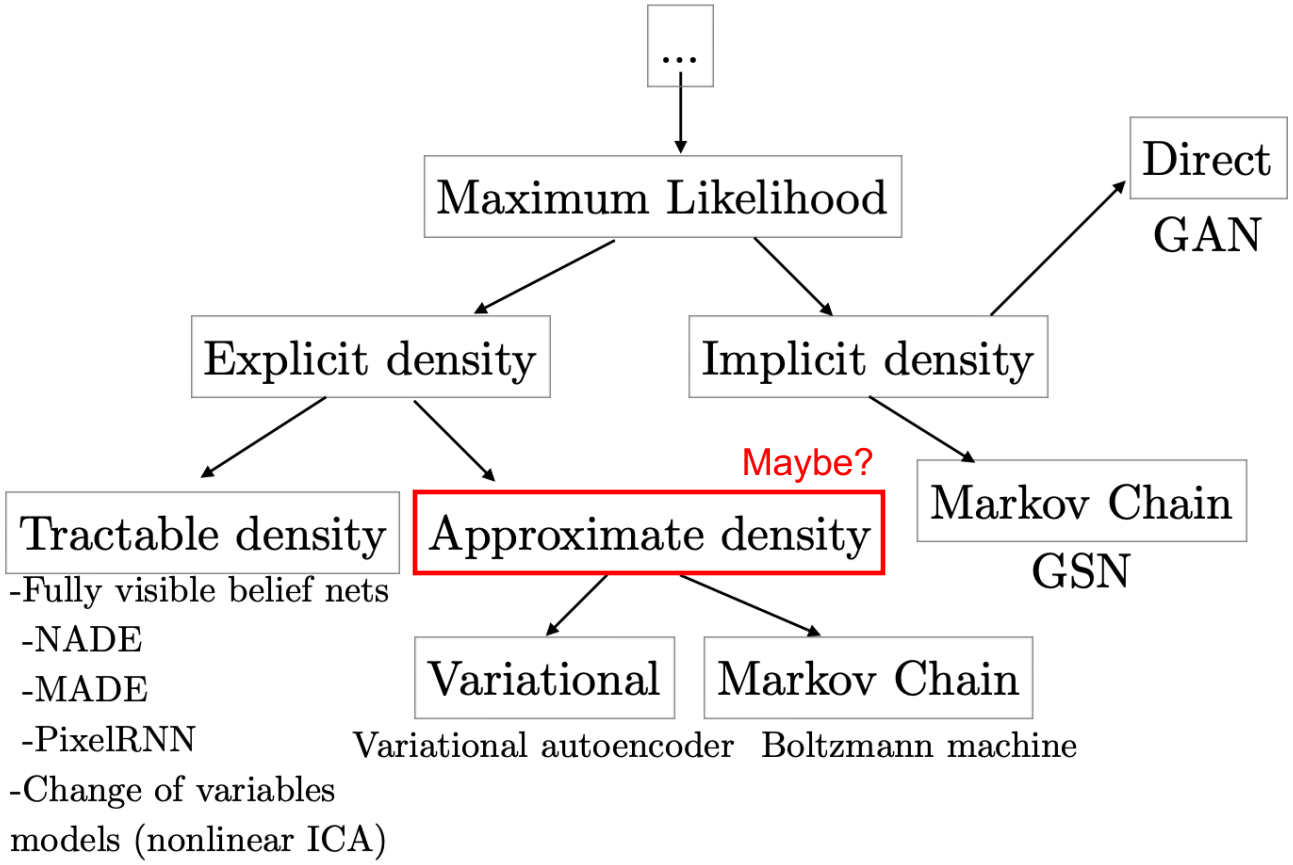
Presented by Minho Park

Contribution

- Score-based generative modeling **(New!)**.
  - Comparable sample quality to GAN.
- Adding noise and annealing the noise level are critical.
  - Annealed Langevin Dynamics.

Generative Models

- GAN taxonomy.
  - Implicit, explicit density estimation.



Explicitly Estimate $p_\theta(x)$

- Explicitly model probabilistic density can be represented as

$$p_\theta(x) = \frac{e^{-f_\theta(x)}}{Z_\theta} \text{ where } Z_\theta = \int e^{-f_\theta(x)} dx\,.$$

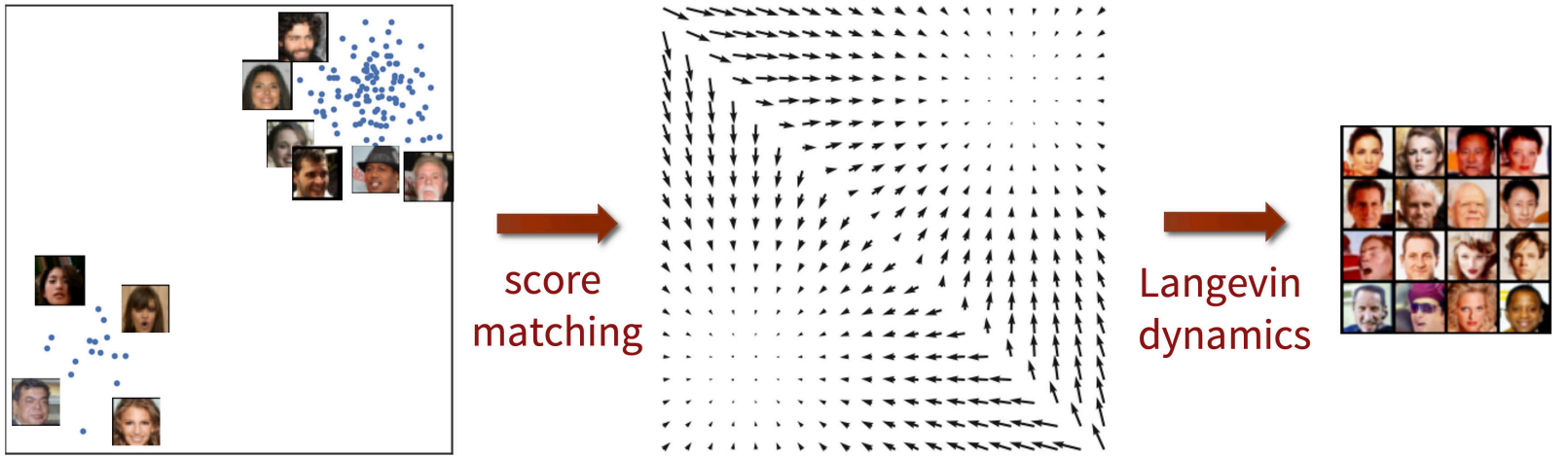
- and empirically maximize

$$\sum_{i=1}^{N} \log p_\theta(x) \quad \text{w.r.t. } \theta.$$

- However, we can not calculate $p_\theta(x)$ because of intractability of Z_θ.

Score Function

- The gradient of a probability density function w.r.t. the input x.
  - Score function forms as vector field.
- $s_\theta(x) \coloneqq \nabla_x \log p_\theta(x) = -\nabla_x \log f_\theta(x) - \nabla_x \log Z_\theta = -\nabla_x \log f_\theta(x)$
  - In statistics, the definition of score function is $\nabla_\theta \log p_\theta(x)$.



Score-based generative modeling with score matching + Langevin dynamics.

Score Estimation

- Given: i.i.d. samples $\{x_1, x_2, \ldots, x_N\} \sim p_{data}(x)$.
- Task: estimate score $\nabla_x \log p_{data}(x)$.
- Score model: a trainable vector-valued function $s_\theta(x)$: $\mathbb{R}^D \to \mathbb{R}^D$.
- Objective: How to compare two vector fields of score?

$$\frac{1}{2}\mathbb{E}_{p_{data}}\left[\frac{1}{2}\|\nabla_x \log p_{data}(x) - s_\theta(x)\|_2^2\right] \quad \text{(Fisher divergence)}$$

- We know that the original distributions will be close if the Fisher divergence is close.
- However, we still can not calculate $\nabla_x \log p_{data}(x)$ which should be used by supervision.

Hyvarinen 05

- Theorem 1.

$$\arg\min_{\theta} \mathbb{E}_{p_{data}}\left[\frac{1}{2}\|\nabla_x \log p_{data}(x) - s_\theta(x)\|_2^2\right] \approx \arg\min_{\theta} \frac{1}{N}\sum_{i=1}^{N}[\frac{1}{2}\|s_\theta(x)\|_2^2 + \mathrm{trace}(\nabla_x s_\theta(x))]$$

- **Remark**
- We don't have to calculate $p_{data}(x)$.
- The minimized case is that

$$s_\theta(x) \to 0, \qquad \nabla_x s_\theta(x) \to -$$

- which means all samples x's are local maxima in probability density function.

Hyvärinen, Aapo, and Peter Dayan. "Estimation of non-normalized statistical models by score matching." *Journal of Machine Learning Research* 6.4 (2005).

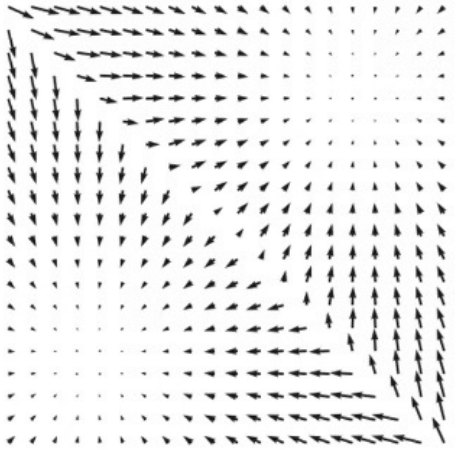
Hyvarinen 05

$$\arg\min_{\theta} \mathbb{E}_{p_{data}}\left[\frac{1}{2}\|\nabla_x \log p_{data}(x) - s_\theta(x)\|_2^2\right] \approx \arg\min_{\theta} \frac{1}{N}\sum_{i=1}^{N}[\frac{1}{2}\|s_\theta(x)\|_2^2 + \text{trace}(\nabla_x s_\theta(x))]$$

- Proof:

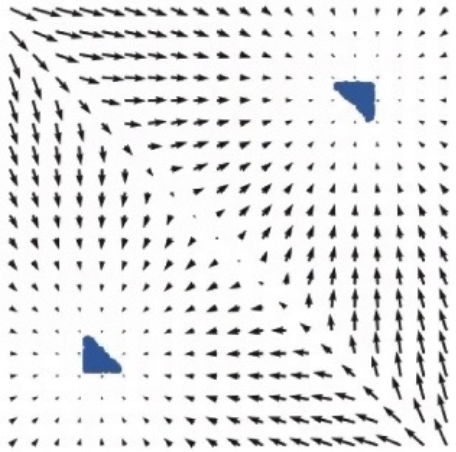
Langevin Dynamics

- A gradient ascent on estimated score function.



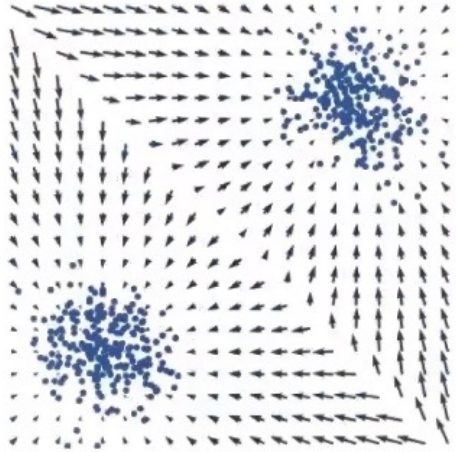
Scores

$$s_\theta(\mathbf{x})$$



Follow the scores

$$\tilde{\mathbf{x}}_{t+1} \leftarrow \tilde{\mathbf{x}}_t + \frac{\epsilon}{2} s_\theta(\tilde{\mathbf{x}}_t)$$



Follow noisy scores: Langevin dynamics

$$\mathbf{z}_t \sim \mathcal{N}(0, I)$$

$$\tilde{\mathbf{x}}_{t+1} \leftarrow \tilde{\mathbf{x}}_t + \frac{\epsilon}{2} s_\theta(\tilde{\mathbf{x}}_t) + \sqrt{\epsilon}\mathbf{z}_t$$

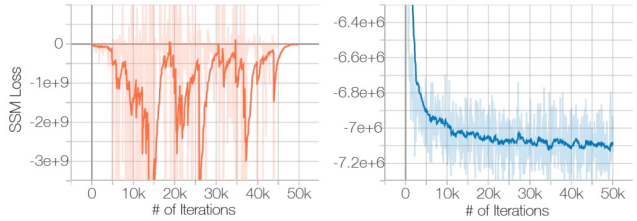
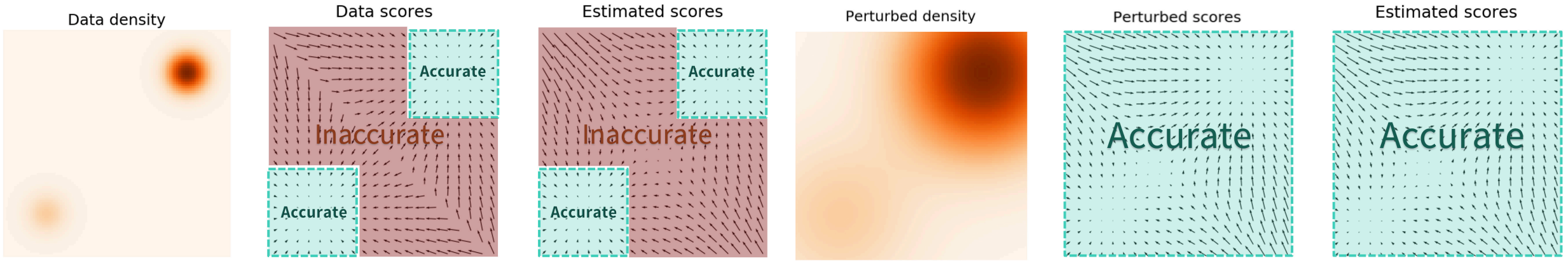


Figure 1: **Left**: Sliced score matching (SSM) loss w.r.t. iterations. No noise is added to data. **Right**: Same but data are perturbed with $\mathcal{N}(0, 0.0001)$.

Annealed Langevin Dynamics

- We estimate $p_{data}(x)$ to our dataset $\{x_1, \dots x_N\} \sim p_{data}(x)$.

$$\arg\min_{\theta} \mathbb{E}_{p_{data}}\left[\frac{1}{2}\|\nabla_x \log p_{data}(x) - s_\theta(x)\|_2^2\right] \approx \arg\min_{\theta} \frac{1}{N}\sum_{i=1}^{N}[\frac{1}{2}\|s_\theta(x)\|_2^2 + \text{trace}(\nabla_x s_\theta(x))]$$

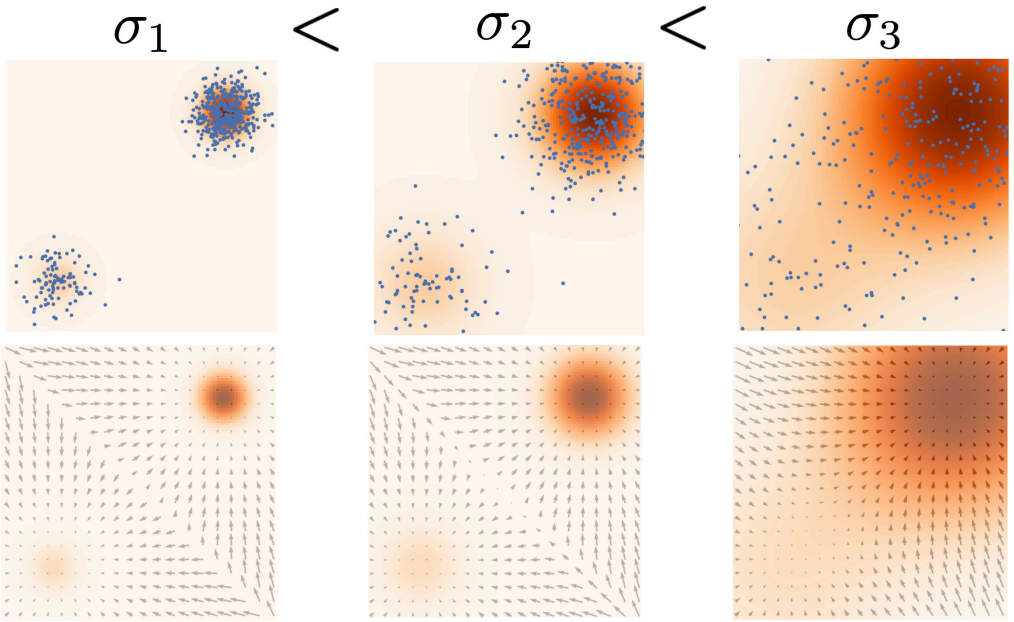


Estimated scores are only accurate in high density regions.

Estimated scores are accurate everywhere for the noise-perturbed data distribution due to reduced low data density regions.

Annealed Langevin Dynamics

- We apply multiple scales of Gaussian noise to perturb the data distribution (first row), and jointly estimate the score functions for all of them (second row).



Practical Recommandations

- Choose $\sigma_1 < \sigma_2 < \cdots < \sigma_L$ as a geometric progression, with σ_1 being sufficiently small and σ_L comparable to the maximum pairwise distance between all training data points.
  - L is typically on the order of hundreds or thousands.
- Parameterize the score-based model $s_\theta(x, i)$ with U-Net skip connections.
- Apply exponential moving average on the weights of the score-based model when used at test time.

Quantitative Results

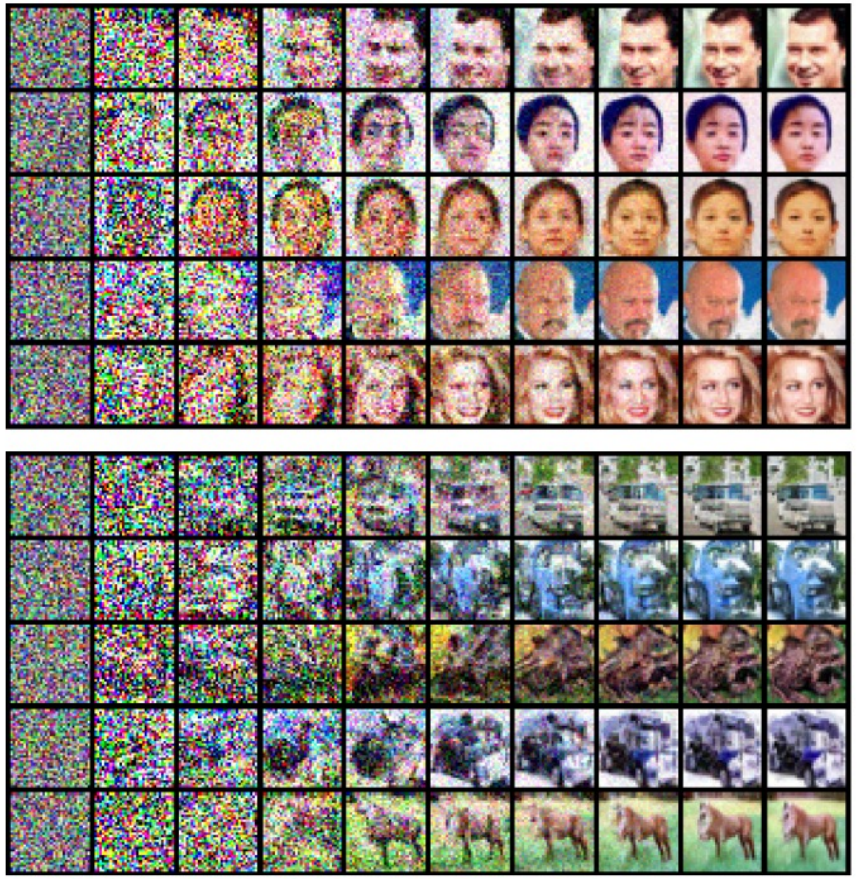


Figure 4: Intermediate samples of annealed Langevin dynamics.

Qualitative Results

Model	Inception	FID
CIFAR-10 Unconditional		
PixelCNN [59]	4.60	65.93
PixelIQN [42]	5.29	49.46
EBM [12]	6.02	40.58
WGAN-GP [18]	7.86 $\pm$ .07	36.4
MoLM [45]	7.90 $\pm$ .10	**18.9**
SNGAN [36]	8.22 $\pm$ .05	21.7
ProgressiveGAN [25]	8.80 $\pm$ .05	-
NCSN (Ours)	**8.87** $\pm$ .12	25.32
CIFAR-10 Conditional		
EBM [12]	8.30	37.9
SNGAN [36]	8.60 $\pm$ .08	25.5
BigGAN [6]	**9.22**	**14.73**

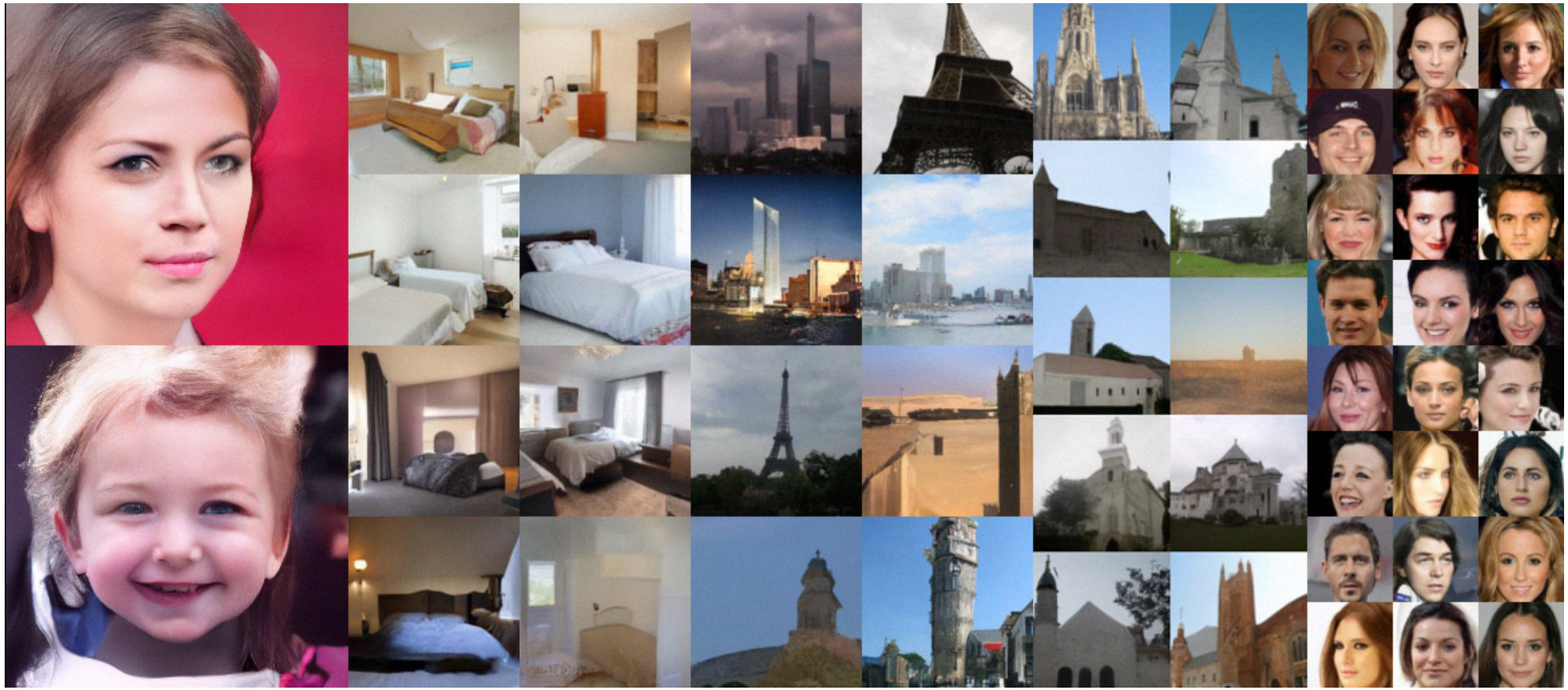
Table 1: Inception and FID scores for CIFAR-10

Quantitative Results

(a) MNIST (b) CelebA (c) CIFAR-10

Figure 5: Uncurated samples on MNIST, CelebA, and CIFAR-10 datasets.

Quantitative Results



Samples from the NCSNv2 model. From left to right:
FFHQ 256x256, LSUN bedroom 128x128, LSUN tower 128x128, LSUN church_outdoor 96x96, and CelebA 64x64.

Quantitative Results

Figure 6: Image inpainting on CelebA (**left**) and CIFAR-10 (**right**). The leftmost column of each figure shows the occluded images, while the rightmost column shows the original images.

Contribution

- Score-based generative modeling **(New!)**.
  - Comparable sample quality to GAN.
- Adding noise and annealing the noise level are critical.
  - Annealed Langevin Dynamics.

Reference

- Stefano Ermon, https://www.youtube.com/watch?v=8TcNXi3A5DI
- Yang Song, https://yang-song.github.io/blog/2021/score/
- Jaejun Yoo, PR12-385 https://www.youtube.com/watch?v=m0sehjymZNU