Accelerating diffusion models for inverse problems through stochastic contraction

Jong Chul Ye

In collaboration with
Hyungjin Chung

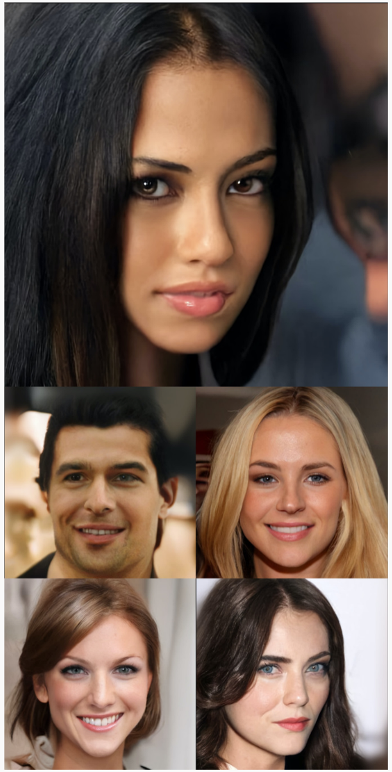
Professor
Graduate School of Artificial Intelligence
KAIST, Korea

KAIST AI
Kim Jaechul Graduate School

Diffusion-based Generative Models

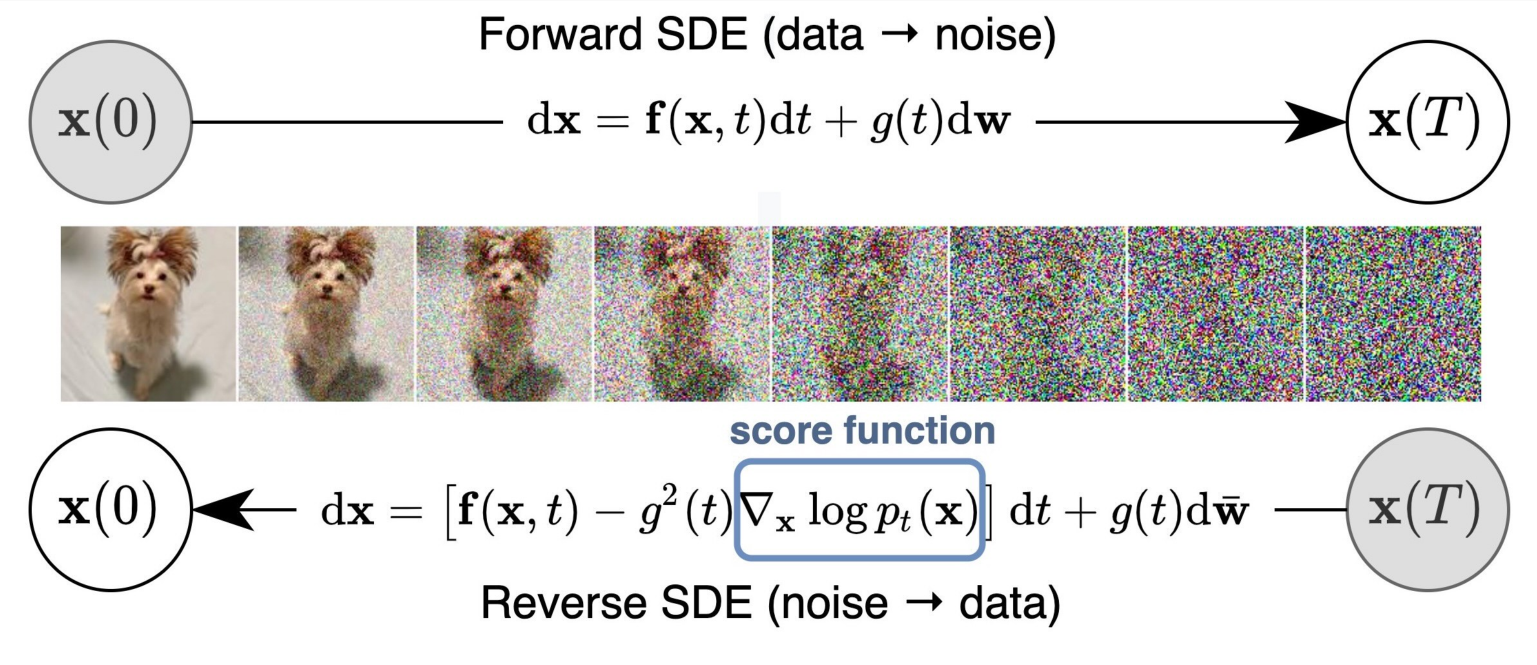
StyleGAN2-ADA
(Karras et al., 2020)

DDPM
(Ho et al., 2020)



Reverse SDE
(Song et al., 2020)

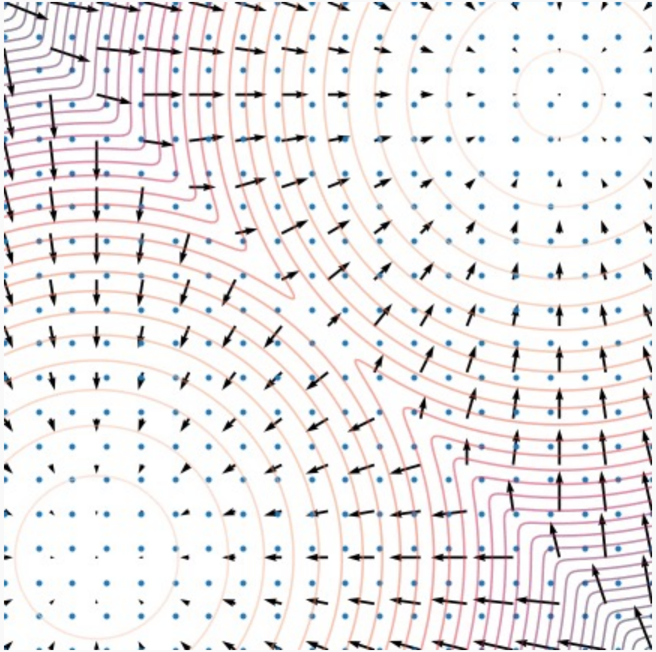
Score-based Generative Models through SDE



Yang Song et al, Generative modeling by estimating gradient of the data distributions, ICLR, 2021

http://yang-song.github.io/blog/2021/score/

Score-based Generative Models through SDE



- Once the score model is trained to optimality,
  - i.e. $s_\theta(\mathbf{x}) \simeq \nabla_{\mathbf{x}} p(\mathbf{x})$
- Use Langevin dynamics to draw samples

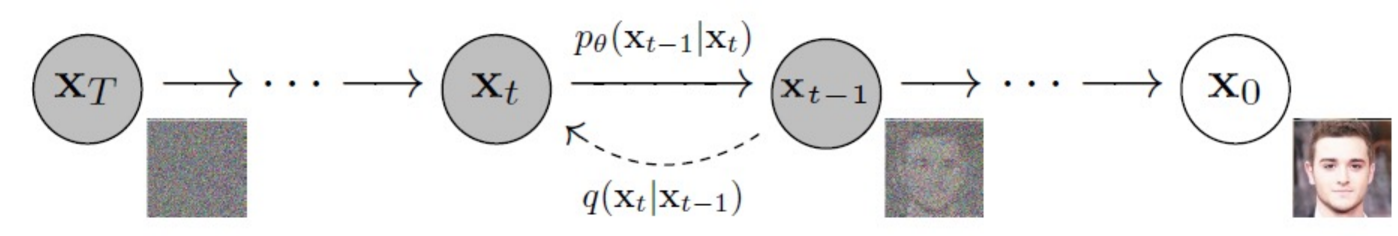
$$\mathbf{x}_{i+1} \leftarrow \mathbf{x}_i + \epsilon \nabla_{\mathbf{x}} \log p(\mathbf{x}) + \sqrt{2\epsilon}\mathbf{z}_i$$

$$i = 0, 1, \ldots, K$$

http://yang-song.github.io/blog/2021/score/

Diffusion Denoising Probabilistic Models (DDPMs)

Ho et al. *NeurIPS*, 2020



- **Train** with **variational lower bound**
- Follow the **reverse markov chain** at **inference**

$$L_{\text{simple}}(\theta) := \mathbb{E}_{t,\mathbf{x}_0,\epsilon}\left[\left\|\epsilon - \epsilon_\theta(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, t)\right\|^2\right]$$

Diffusion Models Beat GANs on Image Synthesis

Dhariwal and Nichol, *NeurIPS*, 2021

Equivalence between the Two Approaches

Song et al. *ICLR* 2022

DDPM

- **Training objective:**

$$L_{\text{simple}}(\theta) := \mathbb{E}_{t,\mathbf{x}_0,\boldsymbol{\epsilon}}\left[\left\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, t)\right\|^2\right]$$

- **Inference:**

SGM

- **Training objective:**

$$\ell(\boldsymbol{\theta};\sigma) \triangleq \frac{1}{2}\mathbb{E}_{p_{\text{data}}(\mathbf{x})}\mathbb{E}_{\tilde{\mathbf{x}}\sim\mathcal{N}(\mathbf{x},\sigma^2 I)}\left[\left\|\mathbf{s}_{\boldsymbol{\theta}}(\tilde{\mathbf{x}},\sigma) + \frac{\tilde{\mathbf{x}}-\mathbf{x}}{\sigma^2}\right\|_2^2\right].$$

- **Inference:**

Reverse Diffusion Through Score-Matching

Noising

$$\mathrm{d}\boldsymbol{x} = \bar{\boldsymbol{f}}(\boldsymbol{x}, t)\, dt + \bar{g}(t)\mathrm{d}\boldsymbol{w}$$

Corresponding reverse SDE

Denoising

$$\mathrm{d}\boldsymbol{x} = [\bar{\boldsymbol{f}}(\boldsymbol{x}, t) - \bar{g}(t)^2 \nabla_x \log p_t(\boldsymbol{x})]dt + \bar{g}(t)\mathrm{d}\boldsymbol{w}$$

$$\simeq [\bar{\boldsymbol{f}}(\boldsymbol{x}, t) - \bar{g}(t)^2 s_\theta(\boldsymbol{x}, t)]dt + \bar{g}(t)\mathrm{d}\boldsymbol{w}$$

- **Solve reverse SDE numerically: Image generation (denoising)**

Song et al. *ICLR* 2022

SCORE-BASED DIFFUSION MODELS FOR INVERSE PROBLEMS

Chung et al, Medical Image Analysis (in revision), 2022

A General Score-based Formula for Inverse Problems

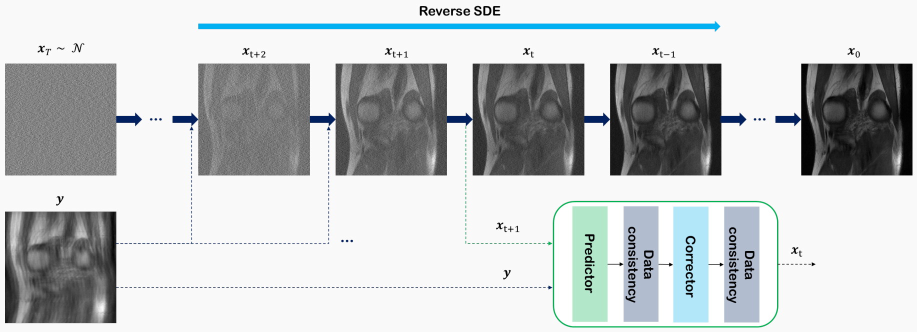
$$\boxed{\min_{\mathbf{x}} \|y - A\mathbf{x}\|^2}$$

$$\boldsymbol{x}_i \leftarrow \boldsymbol{x}_{i+1} + \epsilon_i \boldsymbol{s}_\theta(\boldsymbol{x}_{i+1}, \sigma_{i+1}) + \sqrt{2\epsilon_i}\boldsymbol{z}$$

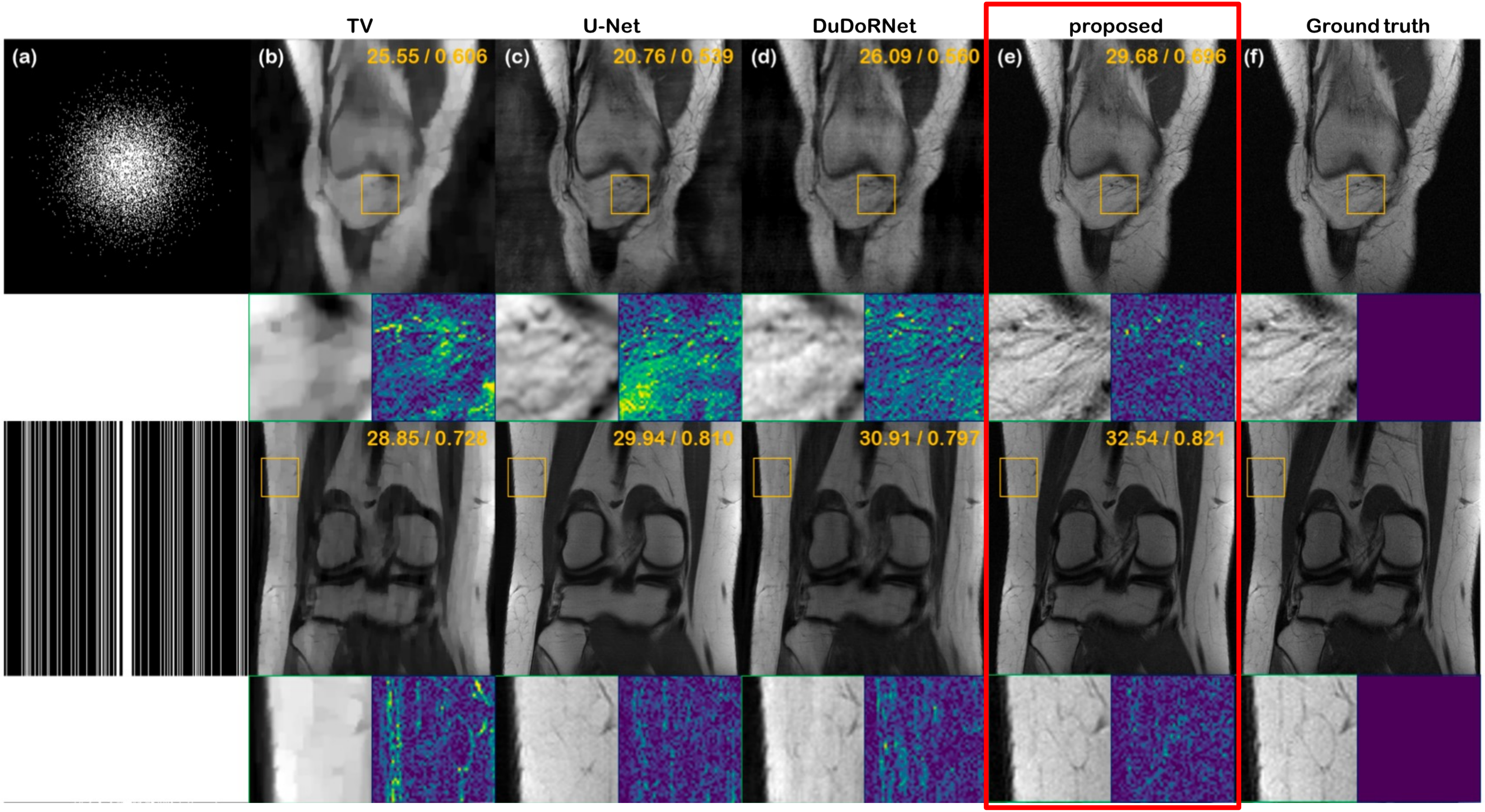
$$\boldsymbol{x}_i \leftarrow \boldsymbol{x}_i + \lambda A^*(y - A\boldsymbol{x}_i),$$

Denoising step (reverse SDE)

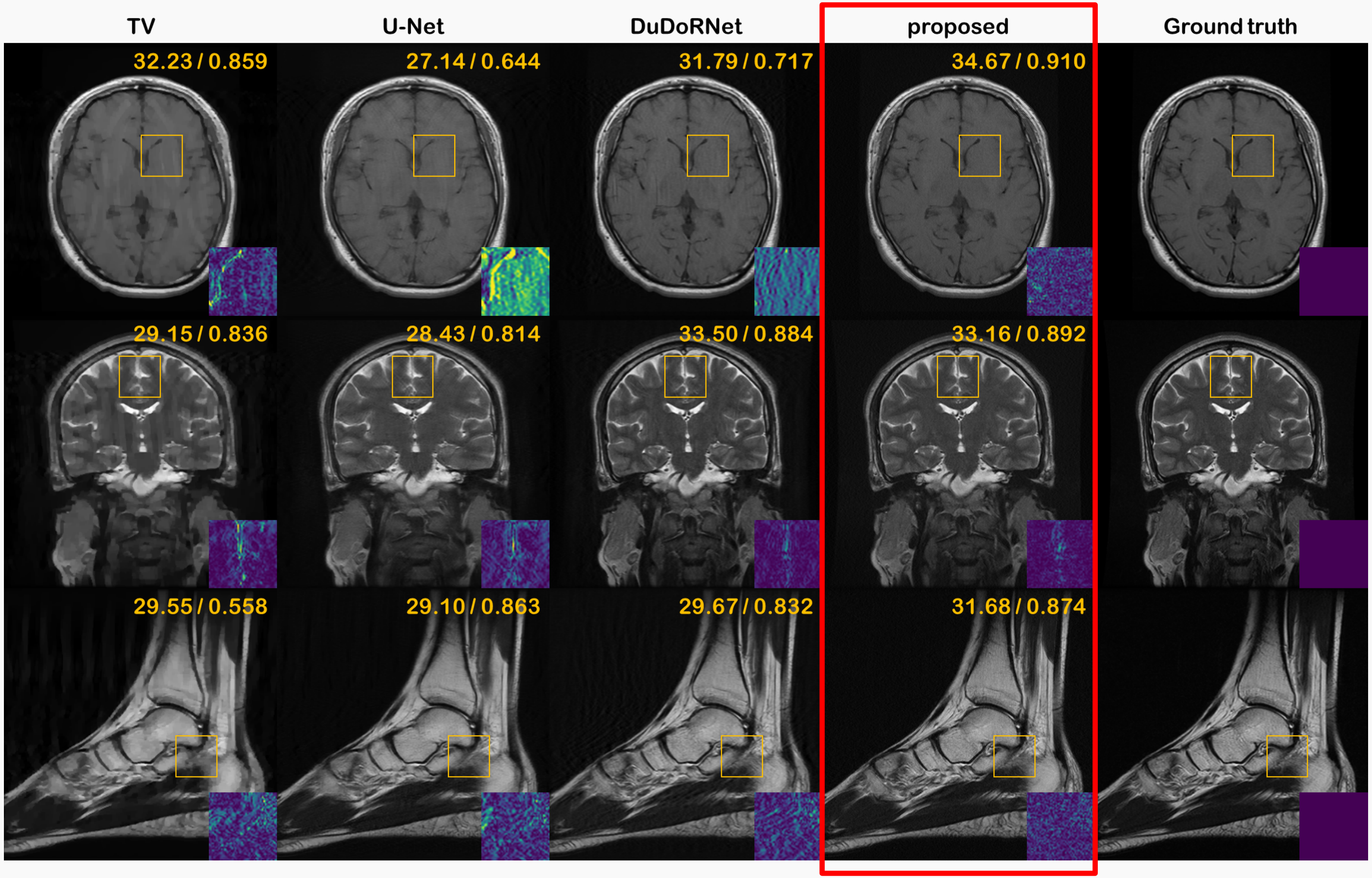
Data consistency step (e.g. GD, POCS)



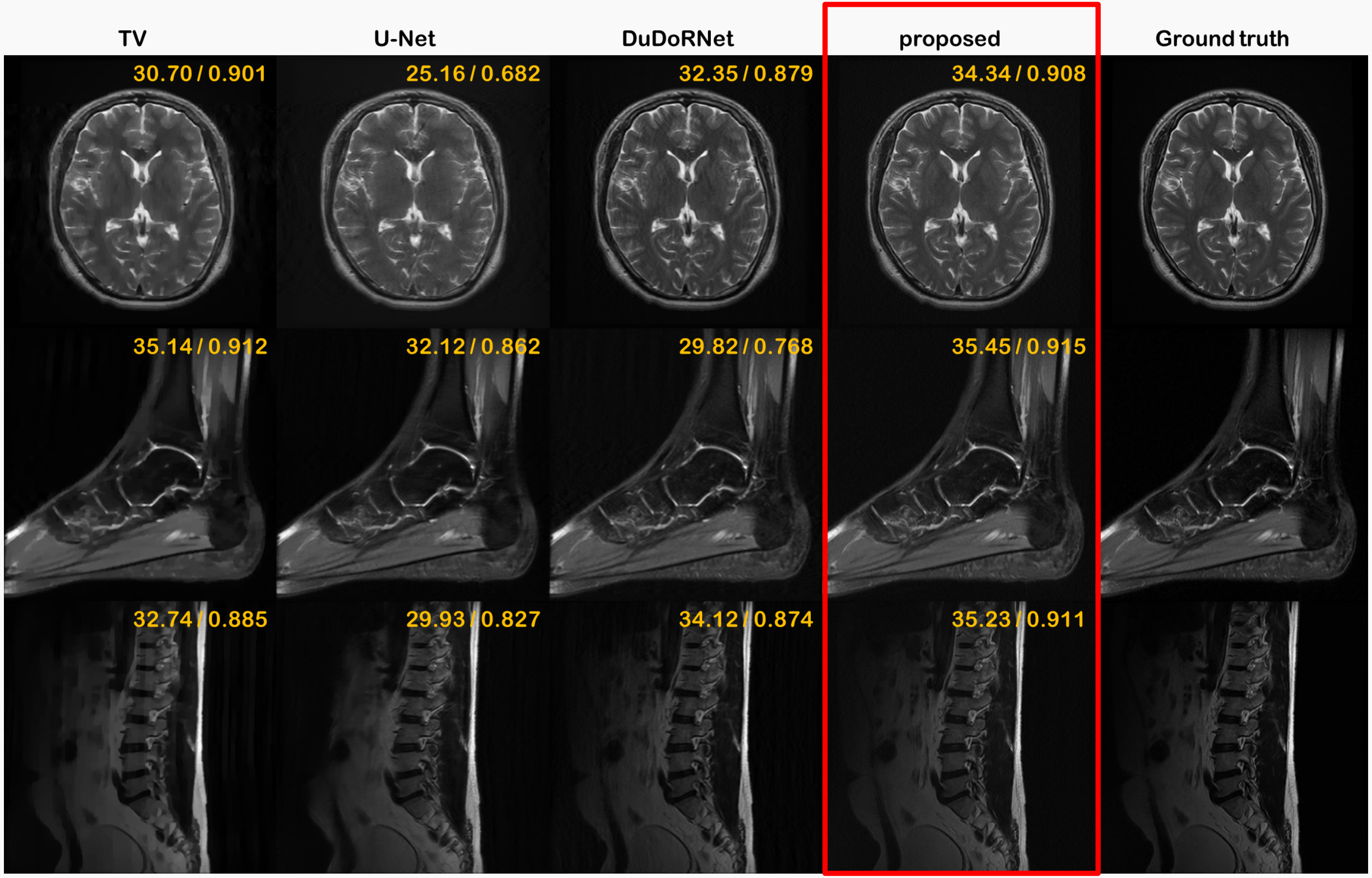
State-of-the-art Performance



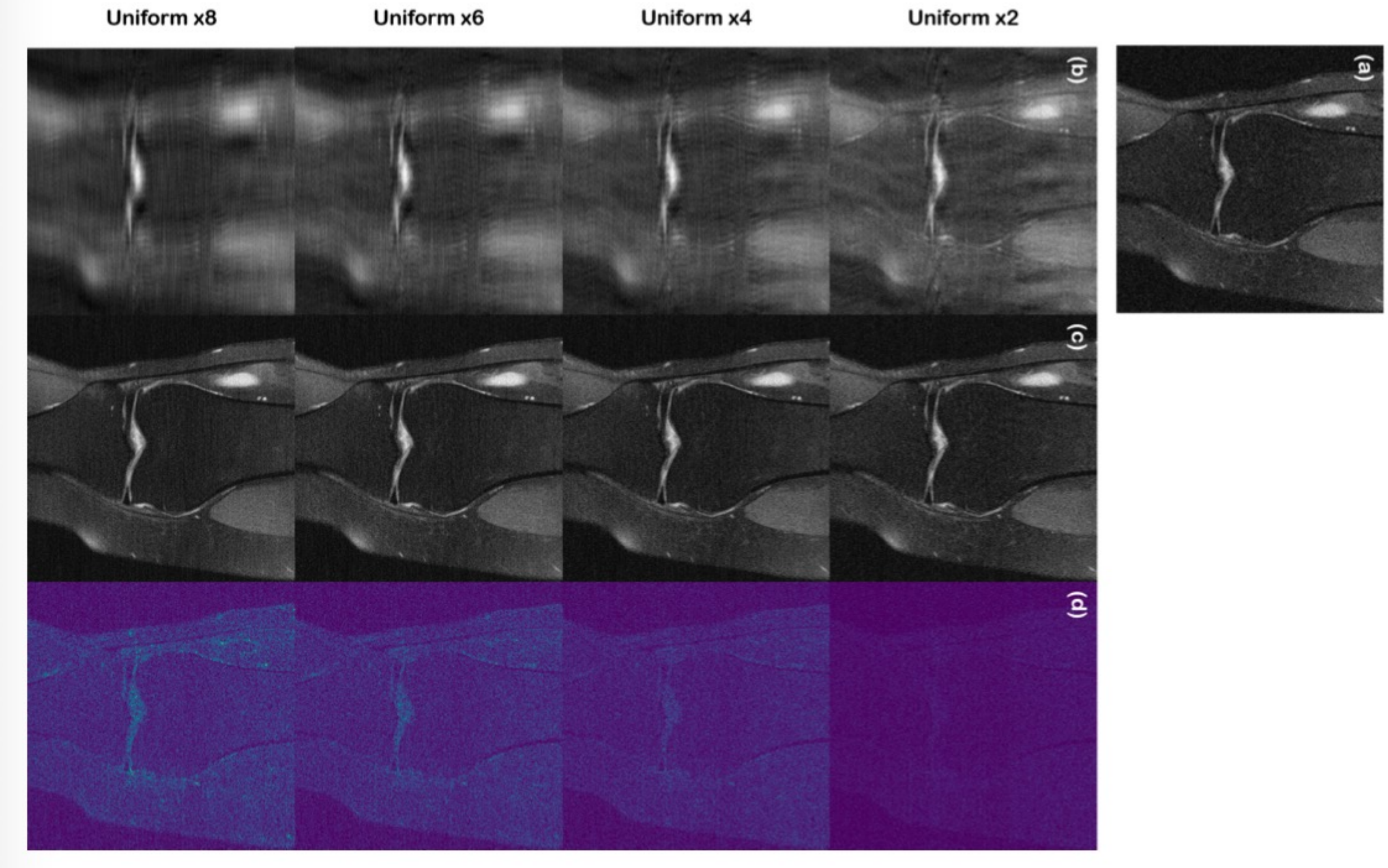
Generalization Capability



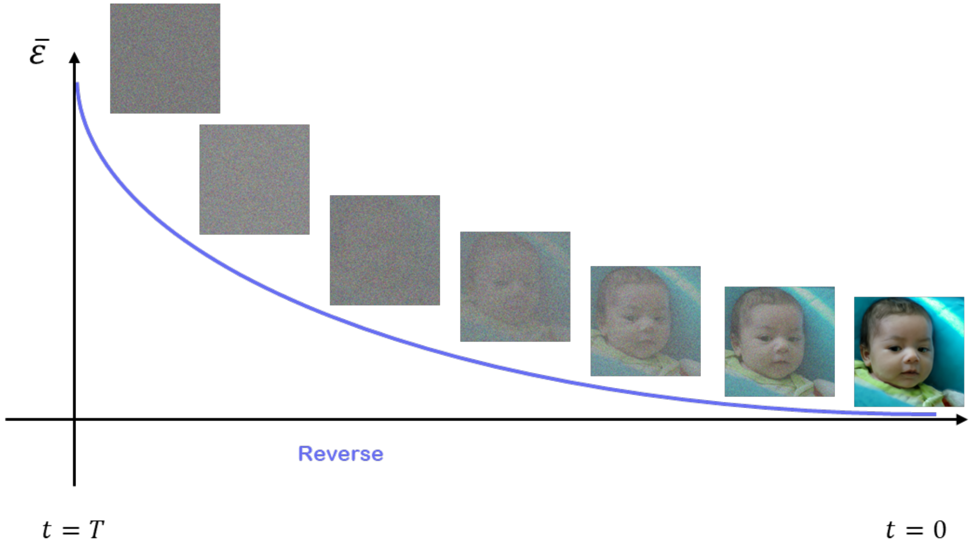
Generalization Capability

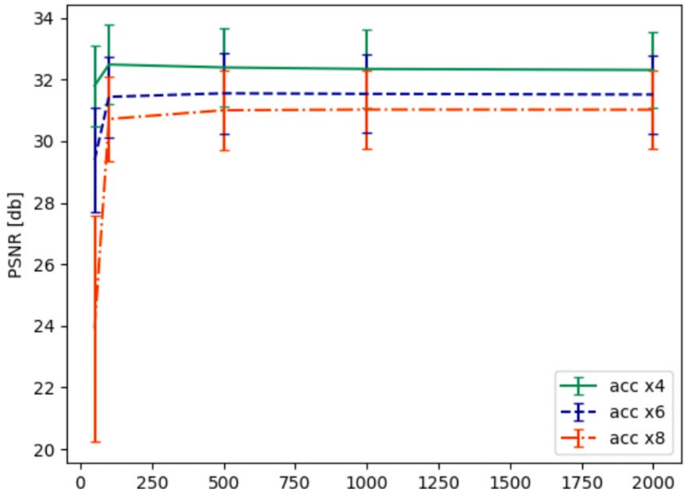


Uncertainty Quantification



Very Slow Convergence

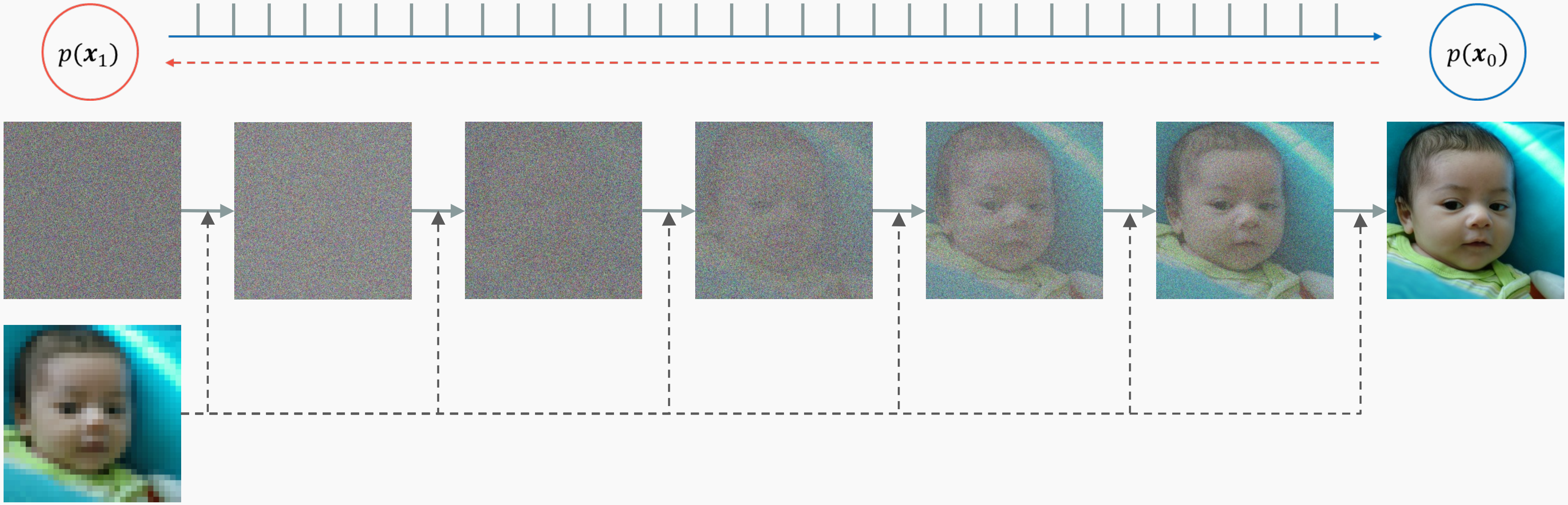




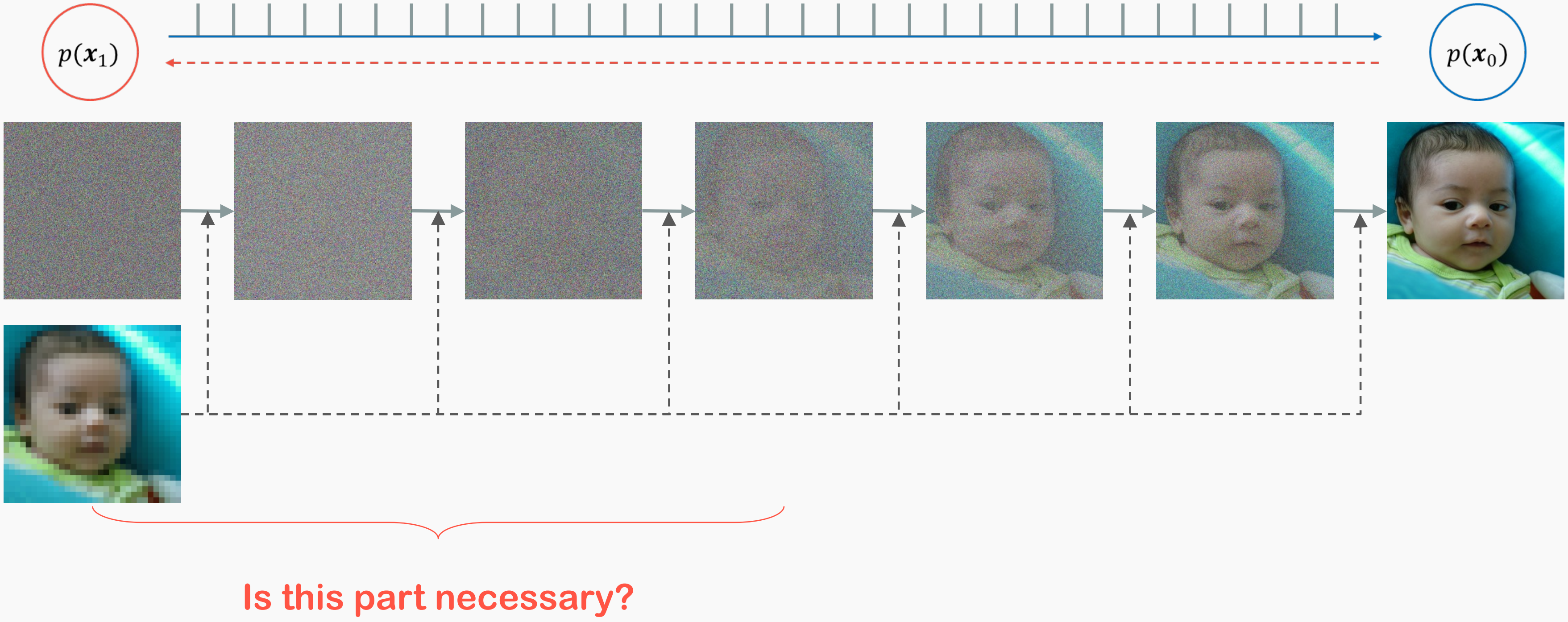
CCDF: COME CLOSER, DIFFUSE FASTER

Chung et al, CVPR, 2022

Intuition: Why use the whole process?

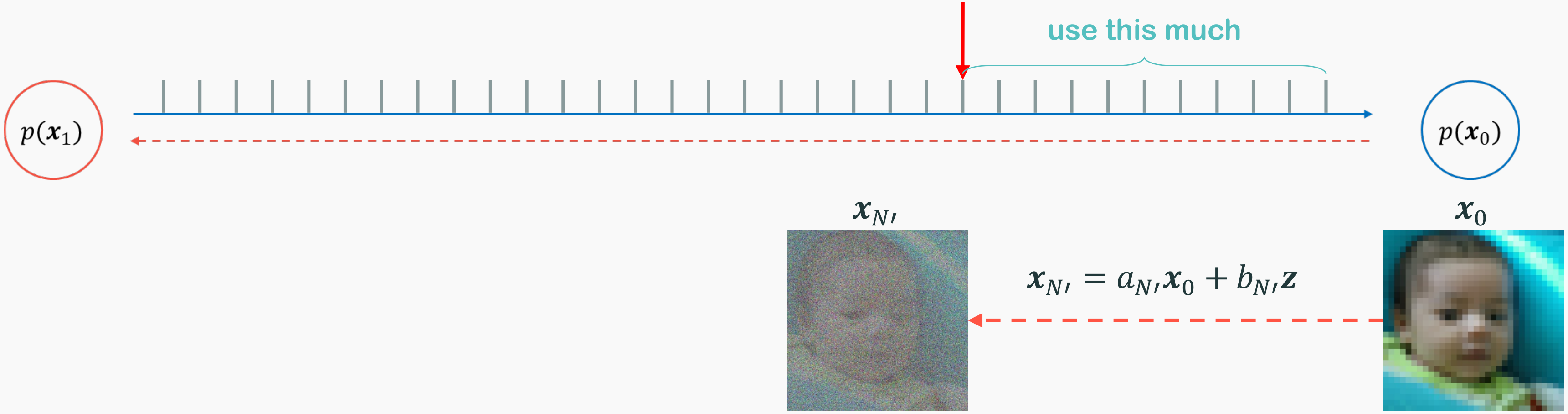


Intuition: Why use the whole process?



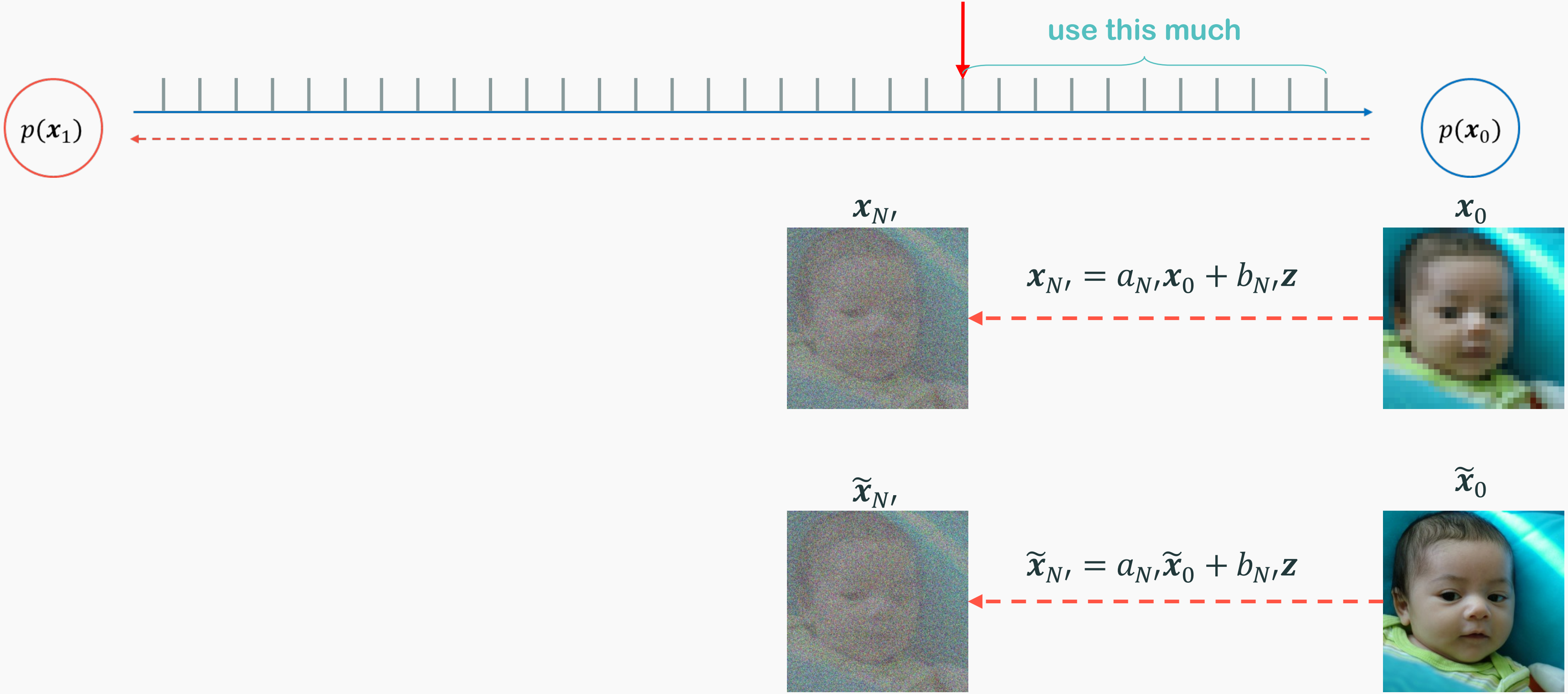
Intuition: Why use the whole process?

- $t_0 = 0.3$
- $N' = t_0 N = 300$



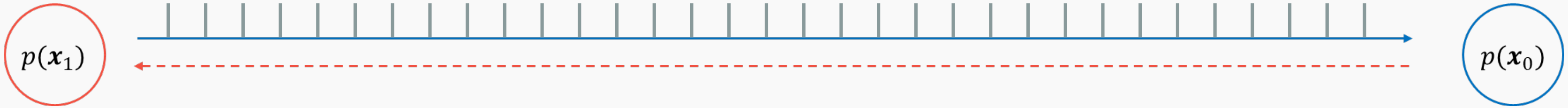
Intuition: Why use the whole process?

- $t_0 = 0.3$
- $N' = t_0 N = 300$



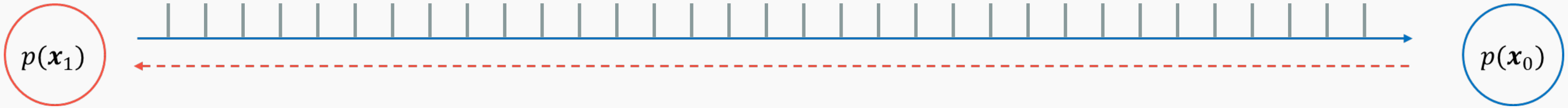
Intuition of CCDF

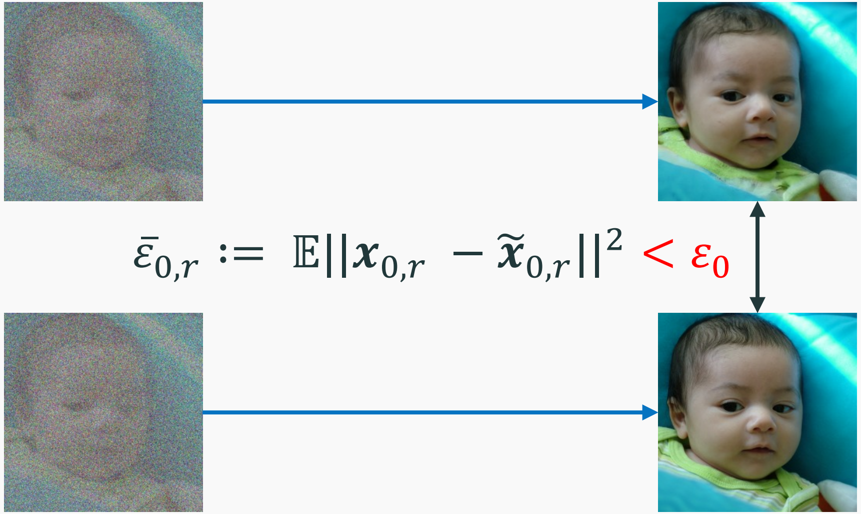
Intuition of CCDF



$$\bar{\varepsilon}_{N'} := \mathbb{E}||\boldsymbol{x}_{N'} - \tilde{\boldsymbol{x}}_{N'}||^2$$

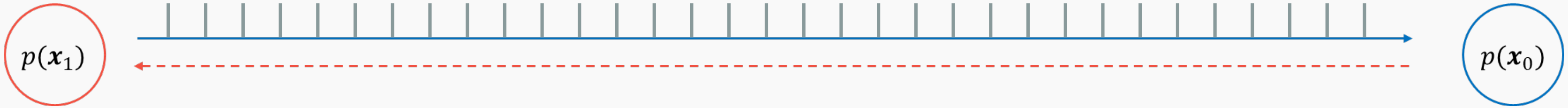
Intuition of CCDF





$$\bar{\varepsilon}_{0,r} := \mathbb{E}||\boldsymbol{x}_{0,r} - \tilde{\boldsymbol{x}}_{0,r}||^2 < \varepsilon_0$$

CCDF: The Algorithm



Algorithm 1 Accelerated Super-resolution / inpainting (VP, Markov)

Require: $\boldsymbol{x}_0, \hat{\boldsymbol{x}}_0, N', \{\alpha_i\}_{i=1}^{N'}, \{\sigma_i\}_{i=1}^{N'}, \boldsymbol{s}_\theta$

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

$\boldsymbol{x}_{N'} \leftarrow \sqrt{\bar{\alpha}_{N'}}\boldsymbol{x}_0 + \sqrt{1-\bar{\alpha}_{N'}}\boldsymbol{z}$ ▷ Forward diffusion

for $i = N'$ to 1 **do** ▷ Reverse diffusion

$\quad \boldsymbol{x}'_{i-1} \leftarrow \frac{1}{\sqrt{\alpha_i}}(\boldsymbol{x}_i + (1-\alpha_i)\boldsymbol{s}_\theta(\boldsymbol{x}_i, i))$

$\quad \boldsymbol{z} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$

$\quad \boldsymbol{x}_{i-1} \leftarrow \boldsymbol{x}'_{i-1} + \sigma_i \boldsymbol{z}$ ▷ Unconditional update

$\quad \boldsymbol{z} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$

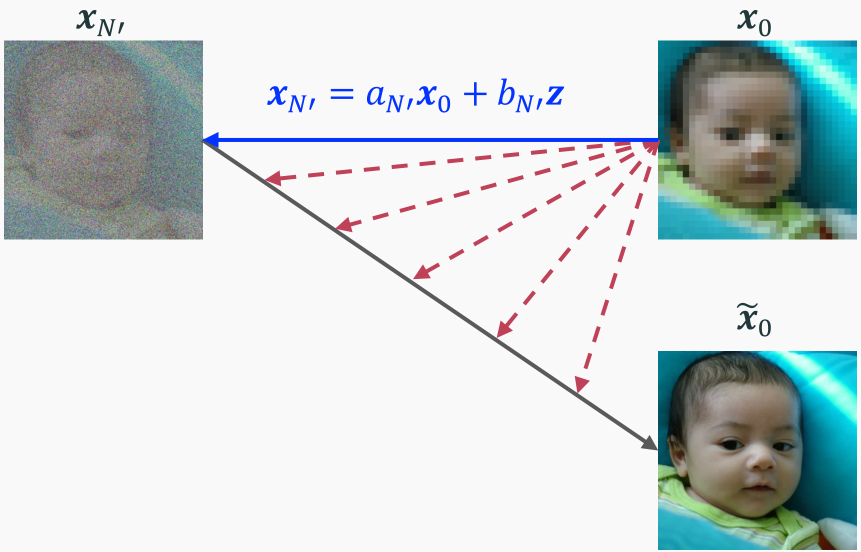
$\quad \hat{\boldsymbol{x}}_i \leftarrow \sqrt{\bar{\alpha}_i}\hat{\boldsymbol{x}}_0 + \sqrt{1-\bar{\alpha}_i}\boldsymbol{z}$

$\quad \boldsymbol{x}_{i-1} = (\boldsymbol{I} - \boldsymbol{P})\boldsymbol{x}_{i-1} + \hat{\boldsymbol{x}}_i$

▷ Measurement consistency

end for

return $\boldsymbol{x}_0$



CCDF: The Algorithm

General form

$$\boldsymbol{x}_{N'} = a_{N'}\boldsymbol{x}_0 + b_{N'}\boldsymbol{z}$$

: 1-step noising

$$\boldsymbol{x}'_{i-1} = \boldsymbol{f}(\boldsymbol{x}_i, i) + \boldsymbol{g}(\boldsymbol{x}_i, i)\boldsymbol{z}_i$$

$$\boldsymbol{x}_{i=1} = \boldsymbol{A}\boldsymbol{x}'_{i-1} + \boldsymbol{b}$$

: Iterative denoising

- **Denoising step** (reverse SDE)
- **Data consistency step** (e.g. GD, POCS)

Constraint

$$\|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{A}\boldsymbol{x}'\| \le \|\boldsymbol{x} - \boldsymbol{x}'\| \ \forall \boldsymbol{x}, \boldsymbol{x}'$$

- **Non-expansive mapping**

Key Idea: Stochastic Contraction

Contraction on $\mathbb{R}^n$

A function $\boldsymbol{f}: \mathbb{R}^n \mapsto \mathbb{R}^n$ contraction mapping ,

if there exists $0 \leq \lambda < 1$ s.t. $\forall\, \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$

$$\|\boldsymbol{f}(\boldsymbol{x}) - \boldsymbol{f}(\boldsymbol{y})\| \leq \lambda \|\boldsymbol{x} - \boldsymbol{y}\|$$

$$\sigma_{\max}\left(\frac{\partial \boldsymbol{f}(\boldsymbol{x})}{\partial \boldsymbol{x}}\right) \leq \lambda < 1$$

Theorem A.1. (Pham et al. 2008)

$$\boldsymbol{x}_{i+1} = f(\boldsymbol{x}_i, i) + g(\boldsymbol{x}_i, i)\boldsymbol{z}$$

- $\boldsymbol{f}$ is contracting with λ
- $\mathrm{Tr}\big(g(\boldsymbol{x}, i)\boldsymbol{I}g(\boldsymbol{x}, i)\big) \leq C \quad \forall \boldsymbol{x}, i$

Then,

$$\mathbb{E}\|x_i - \tilde{x}_i\|^2 \leq \frac{2C}{1 - \lambda^2} + \lambda^{2i}\mathbb{E}\|x_0 - \tilde{x}_0\|^2$$

Reverse SDE is Contracting!

Proof of Theorem 1. (VE-SDE; SMLD)

Forward $\boldsymbol{x}_i = \boldsymbol{x}_0 + \sigma_i z$

Reverse SDE $\boldsymbol{x}'_{i-1} = \underbrace{\boxed{\boldsymbol{x}_i + \left(\sigma_i^2 - \sigma_{i-1}^2\right) s_\theta(\boldsymbol{x}_i, i)}}_{f(x_i, i)} + \sqrt{\sigma_i^2 - \sigma_{i-1}^2}\,\boldsymbol{z}$ **: Stochastically contracting**

Proof.

$$\frac{\partial \boldsymbol{f}^T(\boldsymbol{x}_i, i)}{\partial \boldsymbol{x}_i} = I + \left(\sigma_i^2 - \sigma_{i-1}^2\right) \frac{\partial s_\theta(\boldsymbol{x}_i, i)}{\partial \boldsymbol{x}_i} = \frac{\sigma_{i-1}^2 - \sigma_0^2}{\sigma_i^2 - \sigma_0^2} \boldsymbol{I}$$

$$\lambda = \max_{i \in [N']} \frac{\sigma_{i-1}^2 - \sigma_0^2}{\sigma_i^2 - \sigma_0^2} < 1$$

$$C = \max_{i \in [N']} \sigma_i^2 - \sigma_{i-1}^2$$

Non-expansiveness is Sufficient!

Corollary 1.

$$\boldsymbol{x}'_{i+1} = f(\boldsymbol{x}_i, i) + g(\boldsymbol{x}_i, i)\boldsymbol{z}_i$$

$$\boldsymbol{x}_{i+1} = \underbrace{\boldsymbol{A}}_{\text{Non-expansive mapping}}\boldsymbol{x}'_{i+1} + \boldsymbol{b}$$

$$\mathbb{E}\|x_i - \tilde{x}_i\|^2 \le \frac{2C\tau}{1 - \lambda^2} + \lambda^{2i}\mathbb{E}\|x_0 - \tilde{x}_0\|^2$$

$$\tau = \frac{\mathrm{Tr}(\boldsymbol{A}^T\boldsymbol{A})}{n}$$

Proof.

$$\boldsymbol{x}_{i+1} = \underbrace{\boldsymbol{A}f(x_i, i) + \boldsymbol{b}}_{\tilde{\boldsymbol{f}}(\boldsymbol{x}_i, i)} + \sigma(\boldsymbol{x}_i, i)\boldsymbol{A}\boldsymbol{z}_i$$

$$\sigma_{\max}\left(\frac{\partial\tilde{\boldsymbol{f}}(\boldsymbol{x}, i)}{\partial\boldsymbol{x}}\right) \le \sigma_{\max}(A)\sigma_{\max}\left(\frac{\partial\boldsymbol{f}(\boldsymbol{x}, i)}{\partial\boldsymbol{x}}\right) \le \lambda$$

$$\mathrm{Tr}\big(g(\boldsymbol{x}, i)\boldsymbol{A}^T\boldsymbol{A}g(\boldsymbol{x}, i)\big) = g(\boldsymbol{x}, i)^2\mathrm{Tr}(\boldsymbol{A}^T\boldsymbol{A}) = C\tau$$

Theoretical Findings

Theorem 1.

$$\bar{\varepsilon}_{0,r} \le \frac{2C\tau}{1-\lambda^2} + \lambda^{2N'}\,\bar{\varepsilon}_{N'}$$

Error decreases exponentially with reverse diffusion!

$$\lambda = \begin{cases} \max\limits_{i\in[N']} \sqrt{\alpha_i}\left(\frac{1-\bar{\alpha}_{i-1}}{1-\bar{\alpha}_i}\right) & (DDPM) \\ \max\limits_{i\in[N']} \frac{\sigma_{i-1}^2-\sigma_0^2}{\sigma_i^2-\sigma_0^2} & (SMLD) \\ \max\limits_{i\in[N']} \frac{\sigma_{i-1}}{\sigma_i} & (DDIM) \end{cases}$$

$$C = \begin{cases} n(1-\alpha_N) & (DDPM) \\ n \max\limits_{i\in[N']} \sigma_i^2 - \sigma_{i-1}^2 & (SMLD) \\ 0 & (DDIM) \end{cases}$$

$$\tau = \frac{\mathrm{Tr}(\boldsymbol{A}^T\boldsymbol{A})}{n}$$

Theoretical Findings

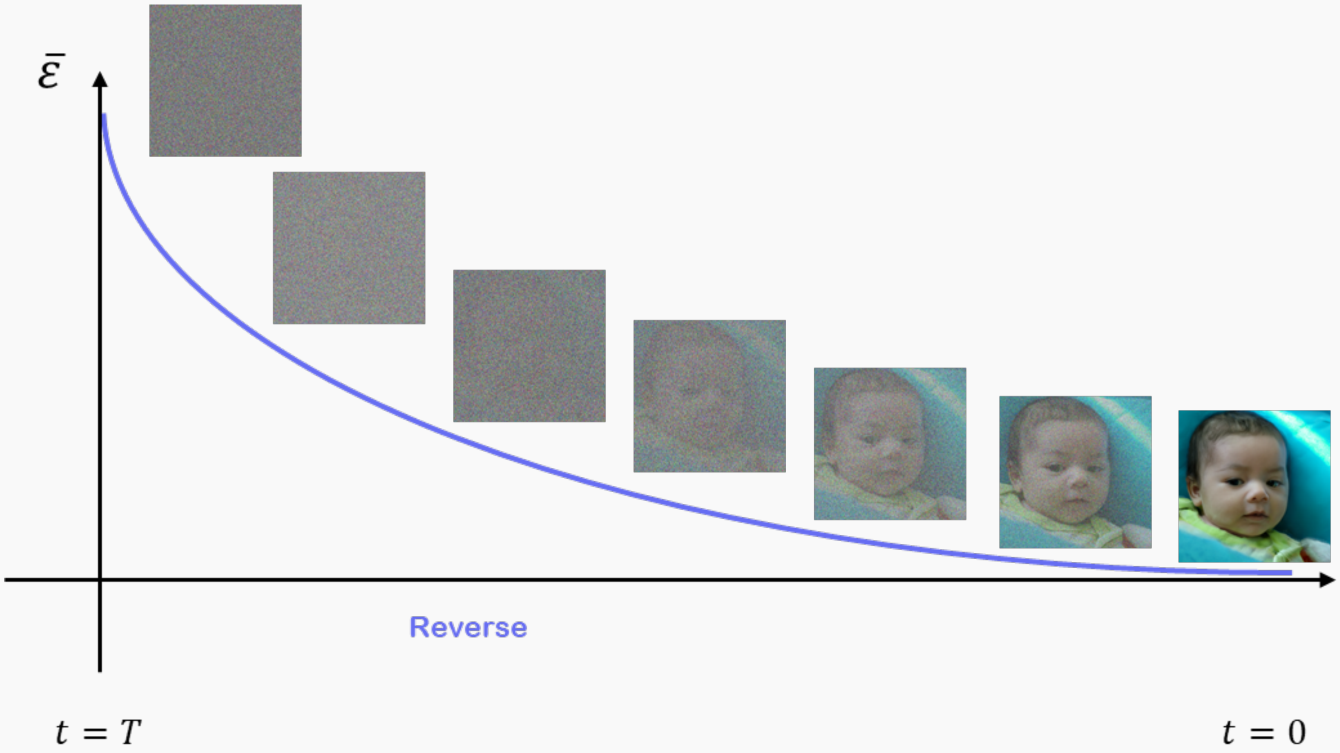
Theorem 2. (shortcut path)

- For any $0 < \mu \leq 1$, there exists a **minimum** N' s.t.

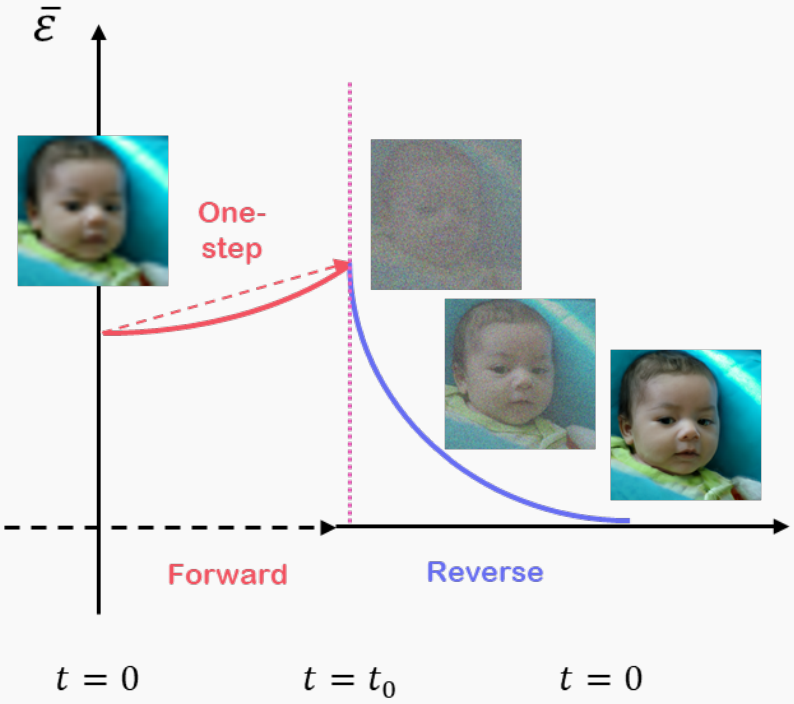
$$\bar{\varepsilon}_{0,r} \leq \mu\varepsilon_0$$

- Optimal N' **decreases** as ε_0 **gets smaller**

Come Closer, Diffuse Faster

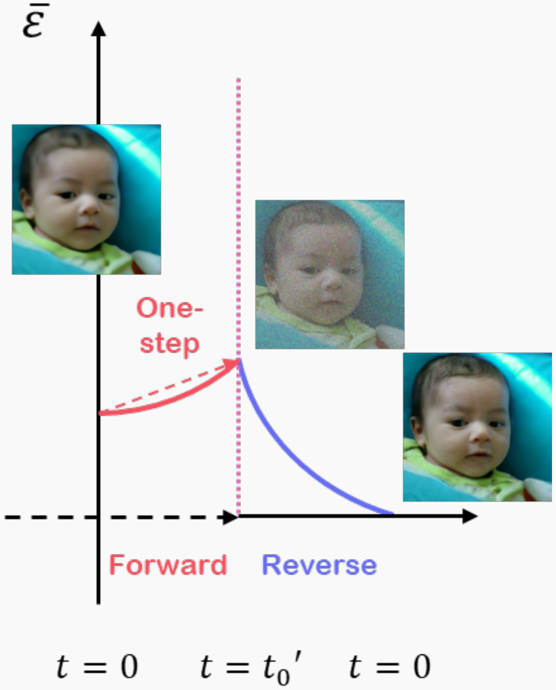


Come Closer, Diffuse Faster

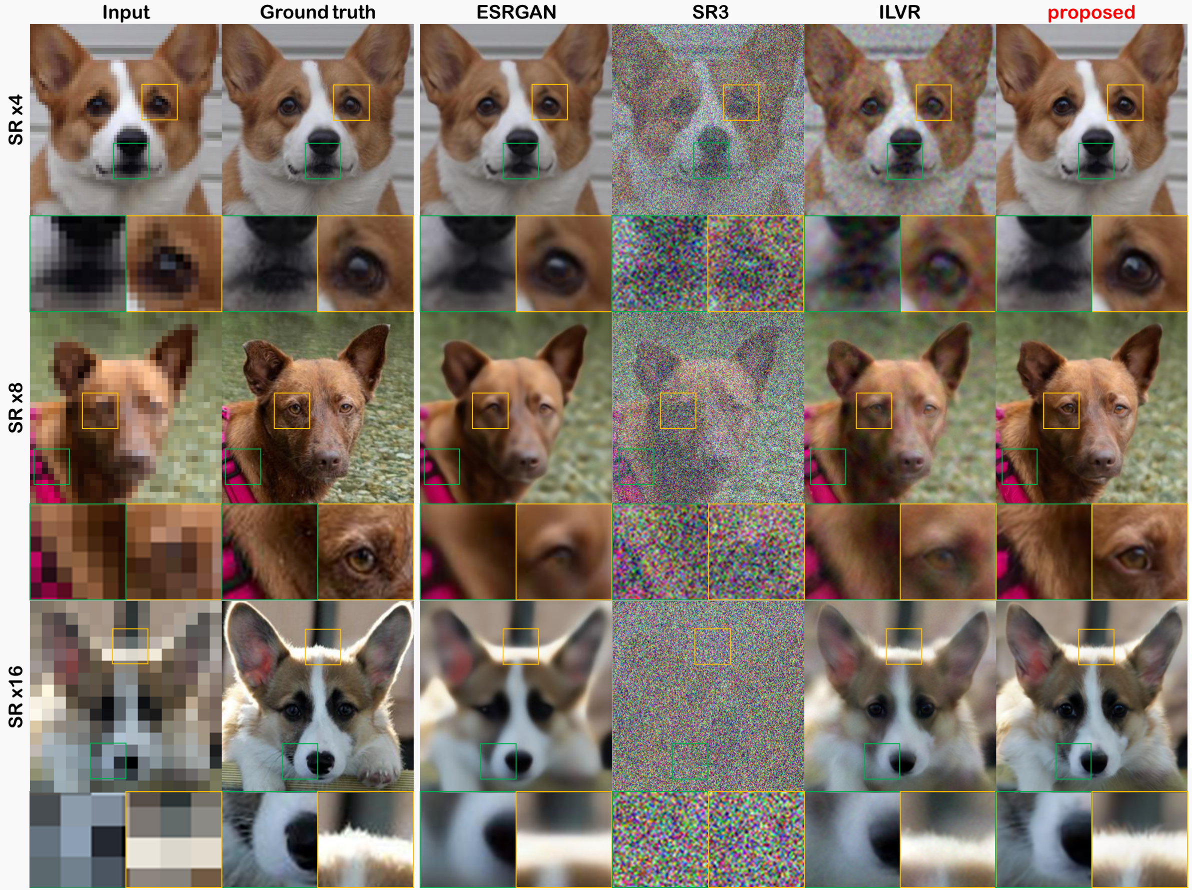


Come Closer, Diffuse Faster

Feed-forward network correction



Experimental Results: SR



20 step diffusion

- **ILVR, SR3**

  $N = 20, \quad t_0 = 1.0$

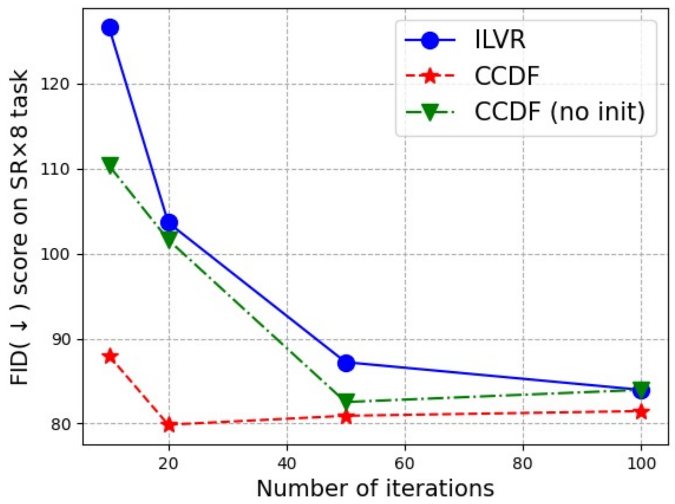
- **proposed**

  $N = 100, \quad t_0 = 0.2$

t_0	0.05	0.1	0.2	0.5	0.75	1.0 [5]
SR $\times 4$	63.90	**60.90**	<u>60.91</u>	64.04	64.14	63.31
SR $\times 8$	85.21	78.13	**75.76**	79.34	79.67	<u>77.34</u>
SR $\times 16$	116.37	101.79	92.59	**88.09**	92.12	<u>88.49</u>

Table 1. FID(↓) scores on FFHQ test set for SR task with $N = 1000$, and varying t_0 values. $t_0 = 1.0$ is the baseline method without any acceleration used in [5]. Numbers in boldface, and underline indicate the best, and the second best scores.

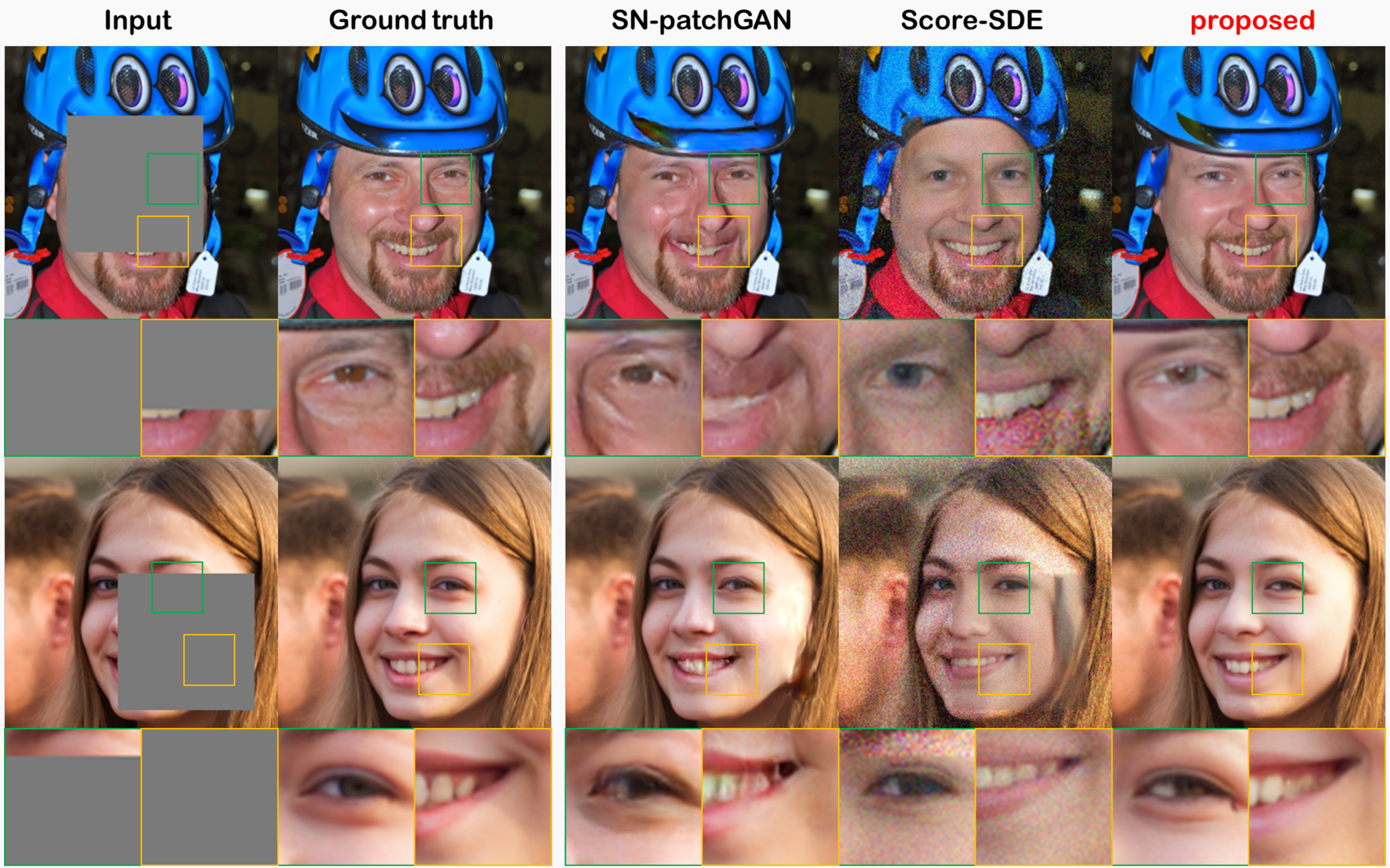
Experimental Results: SR



	SR factor	ESRGAN [36]	SR3* [25]	ILVR [5]	CCDF (ours)
FFHQ	×4	81.14	66.79	63.14	**60.90**
	×8	108.96	80.27	81.85	**75.76**
	×16	143.80	99.46	92.32	**88.39**
AFHQ	×4	24.52	20.68	18.70	**15.53**
	×8	51.84	30.23	34.85	**32.30**
	×16	98.22	60.76	**47.28**	48.77

Table 2. Comparison of FID(↓) scores on FFHQ and AFHQ test set. t_0 values used for the proposed method is $0.1, 0.2, 0.3$ for $\times 4$, $\times 8$, $\times 16$ SR, respectively. Numbers in boldface represent the best results among the row. (*unofficial re-implementation)

Experimental Results: Inpainting



20 step diffusion

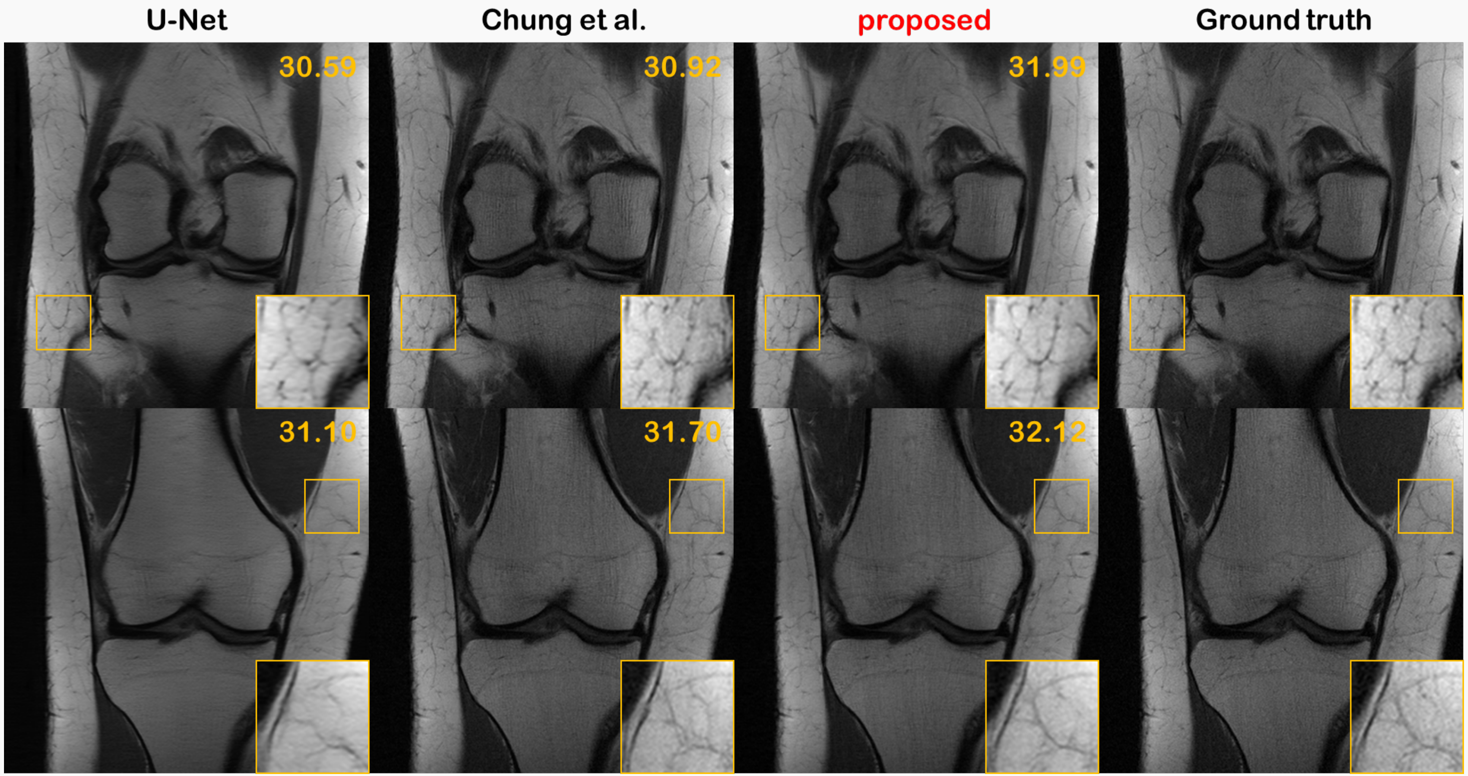
- Score-SDE

  $N = 20, \qquad t_0 = 1.0$

- proposed

  $N = 100, \qquad t_0 = 0.2$

Experimental Results: Fast MRI



20 step diffusion

- Chung et al.

  $N = 1000, \quad t_0 = 1.0$

- proposed

  $N = 1000, \quad t_0 = 0.02$

Summary

- Diffusion models: Exciting new path for solving inverse problems
- Universal solver without knowledge about the problem a priori
- Great generalization capacity
- Acceleration through stochastic contraction theory

Questions?