Come-Closer-Diffuse-Faster: Accelerating Conditional Diffusion Models for Inverse Problems through Stochastic Contraction

Hyungjin Chung[1] Byeongsu Sim[2] Jong Chul Ye[1,2,3]
[1]Bio and Brain Engineering, [2]Mathematical Sciences, [3]Kim Jaechul Graduate School of AI
Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Korea
{hj.chung, byeongsu.s, jong.ye}@kaist.ac.kr

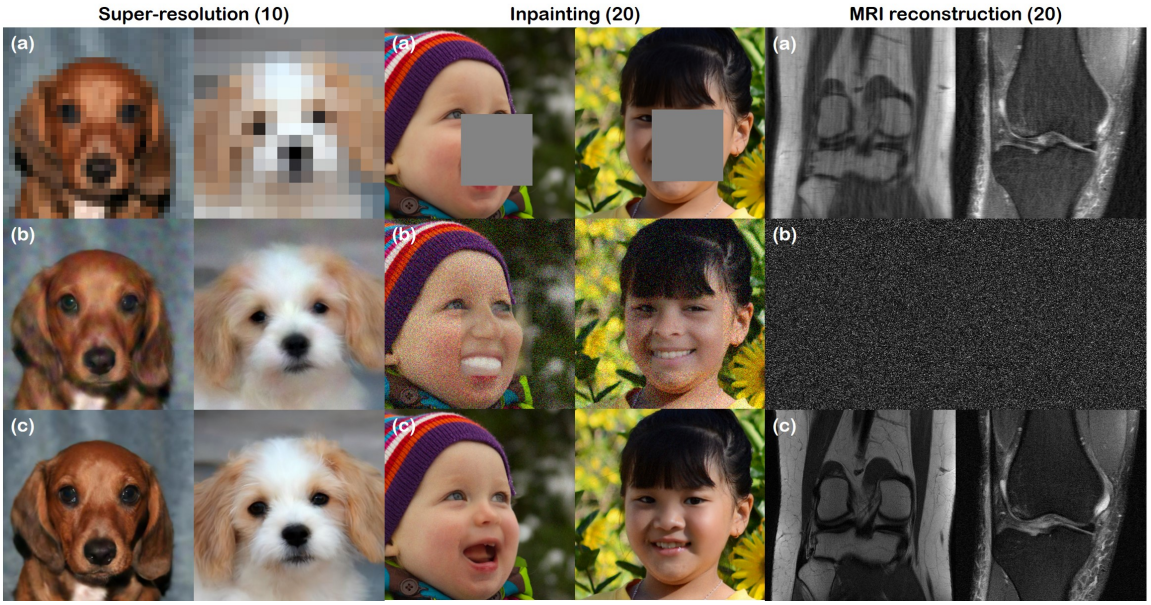

Figure 1. Reconstruction results of three different tasks - super-resolution, inpainting, and MRI reconstruction. Numbers in parenthesis indicate the iteration numbers for reverse diffusion. Proposed method is compared with canonical conditional diffusion models for each task. (a) Corrupted measurement, (b) ILVR [5], score-SDE [34], and score-MRI [6], respectively, for each task. (c) Proposed method.

Abstract

Diffusion models have recently attained significant interest within the community owing to their strong performance as generative models. Furthermore, its application to inverse problems have demonstrated state-of-the-art performance. Unfortunately, diffusion models have a critical downside - they are inherently slow to sample from, needing few thousand steps of iteration to generate images from pure Gaussian noise. In this work, we show that starting from Gaussian noise is unnecessary. Instead, starting from a single forward diffusion with better initialization significantly reduces the number of sampling steps in the reverse conditional diffusion. This phenomenon is formally explained by the contraction theory of the stochastic difference equations like our conditional diffusion strategy - the alternating applications of reverse diffusion followed by a non-expansive data consistency step. The new sampling strategy, dubbed Come-Closer-Diffuse-Faster (CCDF), also reveals a new insight on how the existing feed-forward neural network approaches for inverse problems can be synergistically combined with the diffusion models. Experimental results with super-resolution, image inpainting, and compressed sensing MRI demonstrate that our method can achieve state-of-the-art reconstruction performance at significantly reduced sampling steps.

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No.2019-0-00075, Artificial Intelligence Graduate School Program(KAIST)), and by National Research Foundation(NRF) of Korea grant NRF-2021M3I1A1097938

1. Introduction

Denoising diffusion models [8, 10, 15, 29] and score-based models [31, 32, 34] are new trending classes of gener-

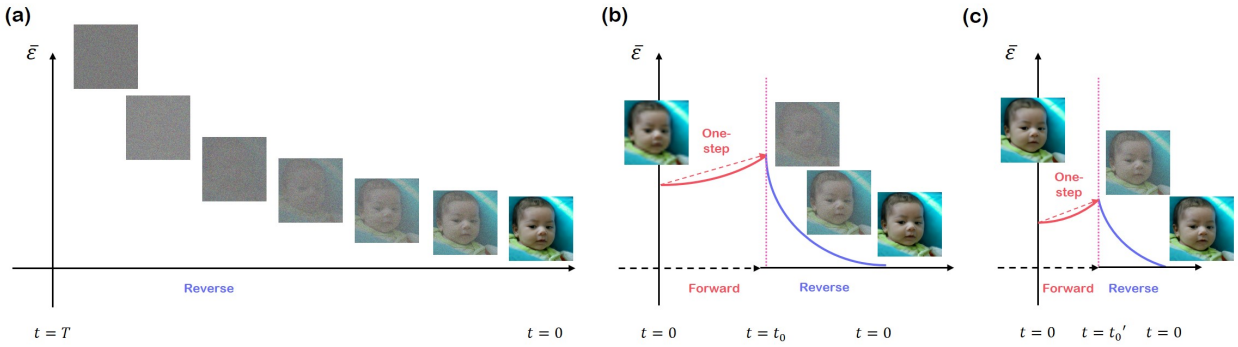

Figure 2. Plot of average error $\bar{\varepsilon}$ vs. time t, using different approaches. (a) Conditional diffusion starts from Gaussian noise $\boldsymbol{x}(t)$ and uses full reverse diffusion. (b) CCDF with vanilla initialization: Corrupted data is forward-diffused with a single step up to $t = t_0$, and reverse diffused. (c) CCDF with NN initialization: Initialization with reconstruction from pre-trained NN lets us use much smaller timestep $t = t_0' < t_0$, and hence faster reverse diffusion.

ative models, which have recently drawn signficant attention amongst the community due to their state-of-the-art performance. Although inspired differently, both classes share very similar aspects, and can be cast as variants of each other [12, 15, 34], thus they are often called *diffusion models*.

In the forward diffusion process, a sampled data point $\boldsymbol{x}$ at time $t = 0$ is perturbed gradually with Gaussian noise until $t = T$, arriving approximately at spherical Gaussian distribution, which is easy to sample from. In the reverse diffusion process, starting from the sampled noise at $t = T$, one uses the trained score function to gradually denoise the data up to $t = 0$, arriving at a high quality data sample.

Interestingly, diffusion models can go beyond unconditional image synthesis, and have been applied to conditional image generation, including super-resolution [5, 17, 25], inpainting [31, 34], MRI reconstruction [6, 13, 33], image translation [5, 19, 28], and so on. One line of works redesign the diffusion model specifically suitable for the task at hand, thereby achieving remarkable performance on the given task [17, 25, 28]. However, they compromise flexibility since the model cannot be used on other tasks. Another line of works, on which we build our method on, keep the training procedure intact, and only modify the inference procedure such that one can sample from a conditional distribution [5, 6, 13, 33, 34]. These methods can be thought of as leveraging the learnt score function as a generative prior of the data distribution, and can be flexibly used across different tasks.

Unfortunately, a critical drawback of diffusion models is that they are very slow to sample from. To address this, for unconditional generative models, many works focused on either constructing deterministic sample paths from the stochastic counterparts [30, 34], searching for the optimal steps to take after the training of the score function [4, 38], or by retraining student networks that can take shortcuts via knowledge distillation [18, 26]. Orthogonal and complementary to these prior works, in this work, we focus on accelerating *conditional* diffusion models by studying the contraction property [21–23] of the reverse diffusion path.

Specifically, our method, which we call *Come-Closer-Diffuse-Faster (CCDF)*, first perturbs the initial estimate via forward diffusion path up to $t_0 < T$, where t_0 denotes the time where the reverse diffusion starts. This forward diffusion comes almost for free, without requiring any passes through the neural network. While the distribution of forward-diffused (noise-added) images increases the estimation errors from the initialization as shown in Fig. 2(b), the key idea of the proposed CCDF is that the reverse conditional diffusion path reduces the error *exponentially* fast thanks to the contraction property of the stochastic difference equation [22, 23]. Therefore, compared to the standard approach that starts the reverse diffusion from Gaussian distribution at $t = T$ (see Fig. 2(a)), the total number of the reverse diffusion step to recover a clean images using CCDF can be significantly reduced. Furthermore, with better initialization, we prove that the number of reverse sampling can be further reduced as shown in Fig. 2(c). This implies that the existing neural-network (NN) based inverse solution can be synergistically combined with diffusion models to yield accurate and fast reconstruction by providing a better initial estimate.

Using extensive experiments across various problems such as super-resolution (SR), inpainting, and MRI reconstruction, we demonstrate that CCDF can significantly accelerate diffusion based models for inverse problems.

2. Background

2.1. Score-based Diffusion Models

We will follow the usual construction of continuous diffusion process $\boldsymbol{x}(t), t \in [0, T]$ with $\boldsymbol{x}(t) \in \mathbb{R}^d$ [34]. Con-

cretely, we want $\boldsymbol{x}(0) \sim p_0(\boldsymbol{x})$, where $p_0 = p_{\text{data}}$, and $\boldsymbol{x}(T) \sim p_T$, where p_T is a tractable distribution that we can sample from. Consider the following Itô stochastic differential equation:

$$d\boldsymbol{x} = \bar{\boldsymbol{f}}(\boldsymbol{x}, t)dt + \bar{g}(t)d\boldsymbol{w}, \tag{1}$$

where $\bar{\boldsymbol{f}} : \mathbb{R}^d \mapsto \mathbb{R}^d$ is the drift coefficient of $\boldsymbol{x}(t)$, $\bar{g} : \mathbb{R} \mapsto \mathbb{R}$ is the diffusion coefficient coupled with the standard d-dimensional Wiener process $\boldsymbol{w} \in \mathbb{R}^d$. By carefully choosing $\bar{\boldsymbol{f}}, \bar{g}$, one can achieve spherical Gaussian distribution as $t \to T$. In particular, when $\bar{\boldsymbol{f}}(\boldsymbol{x}, t)$ is an affine function, then the perturbation kernel $p_{0t}(\boldsymbol{x}(t)|\boldsymbol{x}(0))$ is always Gaussian, where the parameters can be calculated in closed-form. Hence, perturbing the data with the perturbation kernel $p_{0t}(\boldsymbol{x}(t)|\boldsymbol{x}(0))$ comes almost for free, without requiring any passes through the neural network.

For the given forward SDE in (1), there exists a reverse-time SDE running backwards [12, 34]:

$$d\boldsymbol{x} = [\bar{\boldsymbol{f}}(\boldsymbol{x}, t) - \bar{g}(t)^2 \underbrace{\nabla_{\boldsymbol{x}} \log p_t(\boldsymbol{x})}_{\text{score function}}]dt + \bar{g}(t)d\bar{\boldsymbol{w}} \tag{2}$$

where dt is the infinitesimal *negative* time step, and $\bar{\boldsymbol{w}}$ is the Brownian motion running backwards.

Interestingly, one can train a neural network to approximate the actual score function via score matching [31, 34] to estimate $\boldsymbol{s}_\theta(\boldsymbol{x}, t) \simeq \nabla_{\boldsymbol{x}} \log p_t(\boldsymbol{x})$, and plug it into (2) to numerically solve the reverse-SDE [34]. Furthermore, to circumvent technical difficulties, de-noising score matching is typically used where $\nabla_{\boldsymbol{x}} \log p_t(\boldsymbol{x})$ is replaced with $\nabla_{\boldsymbol{x}} \log p_{0t}(\boldsymbol{x}(t)|\boldsymbol{x}(0))$.

2.2. Discrete Forms of SDEs

In this paper, we make use of two different SDEs: variance preserving (VP) SDE, and variance exploding (VE) SDE [34]. First, by choosing

$$\bar{\boldsymbol{f}}(\boldsymbol{x}, t) = -\frac{1}{2}\beta(t)\boldsymbol{x}, \quad \bar{g}(t) = \sqrt{\beta(t)}, \tag{3}$$

where $0 < \beta(t) < 1$ is a monotonically increasing function of noise scale, one achieves the variance preserving (VP)-SDE [10]. On the other hand, variance exploding (VE) SDEs choose

$$\bar{\boldsymbol{f}} = \boldsymbol{0}, \quad \bar{g} = \sqrt{\frac{d[\sigma^2(t)]}{dt}}, \tag{4}$$

where $\sigma(t) > 0$ is again a monotonically increasing function, typically chosen to be a geometric series [31, 34].

For the discrete diffusion models, we assume we have N discretizations which are linearly distributed across $t \in [0, T]$. Then, VP-SDE can be seen as the continuous version of DDPM [15, 34]. Specifically, in DDPM, the forward diffusion is performed as

$$\boldsymbol{x}_i = \sqrt{\bar{\alpha}_i}\boldsymbol{x}_0 + \sqrt{1 - \bar{\alpha}_i}\boldsymbol{z} \tag{5}$$

where $\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$ and $\bar{\alpha}_i = \prod_{j=1}^{i-1} \alpha_j$ for $\alpha_i = 1 - \beta_i$ with monotonically increasing noise schedule $\beta_1, \beta_2, \ldots, \beta_N \in (0, 1)$. The associated reverse diffusion step is

$$\boldsymbol{x}_{i-1} = \frac{1}{\sqrt{\alpha_i}}\Big(\boldsymbol{x}_i + (1 - \alpha_i)\boldsymbol{s}_\theta(\boldsymbol{x}_i, i)\Big) + \sqrt{\sigma_i}\boldsymbol{z}, \tag{6}$$

where $\boldsymbol{s}_\theta(\boldsymbol{x}_i, i)$ is a discrete score function that matches $\nabla_{\boldsymbol{x}_i} \log p_{0i}(\boldsymbol{x}_i|\boldsymbol{x}_0)$. Further, the noise term σ_i can be fixed to $\sigma_i = 1 - \alpha_i$ [10], or set to a learnable parameter [8, 20].

For DDPM, denoising diffusion implicit model (DDIM) establishes the current state-of-the-art among the acceleration methods. Unlike DDPM, DDIM has no additive noise term during the reverse diffusion, allowing less iterations for competitive sample quality. Specifically, the reverse diffusion step is given as:

$$\begin{aligned}\boldsymbol{x}_{i-1} = \sqrt{\bar{\alpha}_{i-1}}\left(\frac{\boldsymbol{x}_i - \sqrt{1 - \bar{\alpha}_i}\boldsymbol{z}_\theta(\boldsymbol{x}_i, i)}{\sqrt{\bar{\alpha}_i}}\right) \\ + \sqrt{1 - \bar{\alpha}_{i-1}}\boldsymbol{z}_\theta(\boldsymbol{x}_i, i)\end{aligned} \tag{7}$$

where

$$\boldsymbol{z}_\theta(\boldsymbol{x}, i) := -\boldsymbol{s}_\theta(\boldsymbol{x}, i)\sqrt{1 - \bar{\alpha}_i}. \tag{8}$$

One can further define

$$\sigma_i = \frac{\sqrt{1 - \bar{\alpha}_i}}{\sqrt{\bar{\alpha}_i}}, \quad \bar{\boldsymbol{x}}_i = \frac{\boldsymbol{x}_i}{\sqrt{\bar{\alpha}_i}}, \tag{9}$$

to express (7) as $\bar{\boldsymbol{x}}_{i-1} = \bar{\boldsymbol{x}}_i + (\sigma_{i-1} - \sigma_i)\boldsymbol{z}_\theta(\boldsymbol{x}_i, i)$.

On the other hand, score matching with Langevin dynamic (SMLD) [31, 32] can be seen as the discrete version of VE-SDE. Specifically, the forward SMLD diffusion step is given by

$$\boldsymbol{x}_i = \boldsymbol{x}_0 + \sigma_i\boldsymbol{z} \tag{10}$$

where $\sigma_i = \sigma_{\min}(\frac{\sigma_{\max}}{\sigma_{\min}})^{\frac{i-1}{N-1}}$, as defined in [34]. The associated reverse diffusion is given by

$$\boldsymbol{x}_{i-1} = \boldsymbol{x}_i + (\sigma_i^2 - \sigma_{i-1}^2)\boldsymbol{s}_\theta(\boldsymbol{x}_i, i) + \sqrt{\sigma_i^2 - \sigma_{i-1}^2}\boldsymbol{z} \tag{11}$$

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

3. Main Contribution

3.1. The CCDF Algorithm

The goal of our CCDF acceleration scheme is to make the reverse diffusion start from $N' := Nt_0 < N$ such that the resulting number of reverse diffusion step can be significantly reduced. For this, our CCDF algorithm is composed of two steps: forward diffusion up to N' with *better* initialization $\boldsymbol{x}_0$, which is followed by a reverse conditional diffusion down to $i = 0$.

Specifically, for a given initial estimate $\boldsymbol{x}_0$, the forward diffusion process can be performed with a single step diffusion as follows:

$$\boldsymbol{x}_{N'} = a_{N'}\boldsymbol{x}_0 + b_{N'}\boldsymbol{z} \tag{12}$$

where $\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$, and $a_{N'}$, $b_{N'}$ for SMLD and DDPM can be computed for each diffusion model using (10) and (5), respectively.

In regard to the conditional difusion, SRDiff [17], SR3 [25] are examples that are trained specifically for SR, with the low-resolution counterparts being encoded or concatenated as the input. However, these approaches attempt to redesign the score function so that one can sample from the conditional distribution, leading to a much complicated formulation.

Instead, here we propose a much simpler but effective conditional diffusion. Specifically, our reverse diffusion uses standard reverse diffusion, alternated with an operation to impose data consistency:

$$\boldsymbol{x}'_{i-1} = \boldsymbol{f}(\boldsymbol{x}_i, i) + g(\boldsymbol{x}_i, i)\boldsymbol{z}_i \tag{13}$$
$$\boldsymbol{x}_{i-1} = \boldsymbol{A}\boldsymbol{x}'_{i-1} + \boldsymbol{b} \tag{14}$$

where the specific forms of $\boldsymbol{f}(\boldsymbol{x}_i, i)$ and $g(\boldsymbol{x}_i, i)$ depend on the type of diffusion models (see Table 1), $\boldsymbol{z}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$, and $\boldsymbol{A}$ is a non-expansive mapping [2]:

$$\|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{A}\boldsymbol{x}'\| \le \|\boldsymbol{x} - \boldsymbol{x}'\|, \quad \forall \boldsymbol{x}, \boldsymbol{x}' \tag{15}$$

In particular, we assume $\boldsymbol{A}$ is linear. For example, one-iteration of the standard gradient descent [13, 24] or projection onto convex sets (POCS) in [9, 11, 27, 35] corresponds to our data consistency step in (14) with (15). See Supplementary Section D for algorithms used for each task.

	$\boldsymbol{f}(\boldsymbol{x}_i, i)$	$g(\boldsymbol{x}_i, i)$
SMLD	$\boldsymbol{x}_i + (\sigma_i^2 - \sigma_{i-1}^2)s_\theta(\boldsymbol{x}_i, i)$	$\sqrt{\sigma_i^2 - \sigma_{i-1}^2}$
DDPM	$\frac{1}{\sqrt{\alpha_i}}(\boldsymbol{x}_i + (1-\alpha_i)s_\theta(\boldsymbol{x}_i, i))$	$\sqrt{1-\alpha_i}$
DDIM	$\sqrt{\alpha_{i-1}}\left(\frac{\boldsymbol{x}_i - \sqrt{1-\bar{\alpha}_i}z_\theta(\boldsymbol{x}_i, i)}{\sqrt{\bar{\alpha}_i}}\right) + \sqrt{1-\bar{\alpha}_{i-1}}z_\theta(\boldsymbol{x}_i, i)$	0

Table 1. Values of $\boldsymbol{f}$, g, and noise schedule of discrete SDEs.

3.2. Fast Convergence Principle of CCDF

Now, we are ready to show why CCDF provides much faster convergence than the standard conditional diffusion models that starts from Gaussian noise. In fact, the key innovation comes from the mathematical findings that while the forward diffusion increases the estimation error, the conditional reverse diffusion decreases it much faster at exponential rate. Accordingly, we can find a "sweet spot" N' such that the forward diffusion up to N' followed by reverse diffusion can significantly reduces the estimation error of the initial estimate $\boldsymbol{x}_0$. This fast convergence principle is shown in the following theorems, whose proofs can be found in Supplementary Materials. First, the following lemma is a simple consequence of independency of Gaussian noises.

Lemma 1. *Let $\tilde{\boldsymbol{x}}_0 \in \mathbb{R}^n$ and $\boldsymbol{x}_0 \in \mathbb{R}^n$ be the ground-truth clean image and its initial estimate, respectively, and the initial estimation error is denoted by $\varepsilon_0 = \|\boldsymbol{x}_0 - \tilde{\boldsymbol{x}}_0\|^2$. Suppose, furthermore, that $\boldsymbol{x}_{N'}$ and $\tilde{\boldsymbol{x}}_{N'}$ denote the forward diffusion from $\boldsymbol{x}_0$ and $\tilde{\boldsymbol{x}}_0$, respectively, using (12). Then, the estimation error after the forward diffusion is given by*

$$\begin{aligned}\bar{\varepsilon}_{N'} &:= \mathbb{E}\|\boldsymbol{x}_{N'} - \tilde{\boldsymbol{x}}_{N'}\|^2 \\ &= a_{N'}^2\varepsilon_0 + 2b_{N'}^2 n.\end{aligned} \tag{16}$$

Now, the following theorem, which is a key step of our proof, comes from the stochastic contraction property of the stochastic difference equation [22, 23].

Theorem 1. *Consider the reverse diffusion using (13) and (14). Then, we have*

$$\bar{\varepsilon}_{0,r} \le \frac{2C\tau}{1-\lambda^2} + \lambda^{2N'}\bar{\varepsilon}_{N'} \tag{17}$$

where $\bar{\varepsilon}_{0,r}$ denotes the estimation error between reverse conditional diffusion path down to $i = 0$, and $\tau = \frac{\mathrm{Tr}(\boldsymbol{A}^T\boldsymbol{A})}{n}$. Furthermore, the contraction rate λ and the constant C have the following closed form expression:

$$\lambda = \begin{cases} \max\limits_{i\in[N']} \sqrt{\alpha_i}\left(\frac{1-\bar{\alpha}_{i-1}}{1-\bar{\alpha}_i}\right) & \textit{(DDPM)} \\ \max\limits_{i\in[N']} \frac{\sigma_{i-1}^2 - \sigma_0^2}{\sigma_i^2 - \sigma_0^2} & \textit{(SMLD)} \\ \max\limits_{i\in[N']} \frac{\sigma_{i-1}}{\sigma_i} & \textit{(DDIM)} \end{cases} \tag{18}$$

and

$$C = \begin{cases} n(1-\alpha_N) & \textit{(DDPM)} \\ n \max\limits_{i\in[N']} \sigma_i^2 - \sigma_{i-1}^2 & \textit{(SMLD)} \\ 0 & \textit{(DDIM)} \end{cases} \tag{19}$$

Now we have the main results that shows the existence of the shortcut path for the acceleration.

Theorem 2 (Shortcut path)**.** *For any $0 < \mu \le 1$, there exists a minimum $N'(= t_0 N < N)$ such that $\bar{\varepsilon}_{0,r} \le \mu\varepsilon_0$. Furthermore, N' decreases as ε_0 gets smaller.*

Theorem 1 states that the conditional reverse diffusion is exponentially contracting. Subsequently, Theorem 2 tells us that we can achieve superior results (i.e. tighter bound) with *shorter* sampling path. Hence, it is unnecessary for us to start sampling from N. Rather, we can start from an arbitrary timestep $N' < N$, and still converge faster to the same point that could be achieved when starting the sampling procedure at N. Furthermore, as we have better initialization

such that ε_0 is smaller, then we need smaller reverse diffusion step, achieving much higher acceleration.

For example, we can initialize the corrupted image with a pre-trained neural network G_φ, which has been widely studied across different tasks [16,39,40]. These methods are typically extremely fast to compute, and thus does not introduce additional computational overload. Using this rather simple and fast fix, we observe that we are able to choose smaller values of t_0, endowed with much stabler performance. For example, in the case of MRI reconstruction, we can choose t_0 as small as 0.02, while *outperforming* score-MRI [6] with 50× acceleration.

4. Experiments

4.1. Experimental settings

We test our method on three different tasks: super-resolution, inpainting, and MRI reconstruction. For all methods, we evaluate the qualitative image quality and quantitative metrics as we accelerate the diffusion process by reducing the t_0 values. For the proposed method, we report on the results starting with neural network (NN)-initialized $\boldsymbol{x}_0$ unless specified otherwise.

Dataset. For vision tasks using face images, we use two datasets - FFHQ 256×256, and AFHQ 256×256. For FFHQ, we randomly select 50k images for training, and sample 1k images of test data separately. For AFHQ, we train our model using the images in the `dog` category, which consists of about 5k images. Testing was performed with the held-out validation set of 500 images of the same category. For the MRI reconstruction task, we use the fastMRI knee data, which consists of around 30k 320×320-sized slices of coronal knee scans. Specifically, we use magnitude data given as the key `reconstruction_esc`. We randomly sample 10 volumes from the validation set for testing.

Quantitative metrics. Since it is well known that for high corruption factors, standard metrics such as PSNR/SSIM does not correlate well with the visual quality of the reconstruction [25, 40], we report on the FID score based on `pytorch-fid`[1]. For MRI reconstruction, it is less sound to report on FID; hence, we report on PSNR.

Super-resolution. Experiments were performed across three different levels of SR factor - $\times 4, \times 8, \times 16$. We train a discretized VP-SDE based on IDDPM [20] for each dataset - FFHQ and AFHQ, following the standards. Specific details can be found in Supplementary section D. For the one-step feed forward network corrector, we train the widely-used ESRGAN [37] for each SR factor, using the same neural network architecture that was used to train the score function. We use three methods for comparison - ESRGAN, ILVR, and SR3[2]. We note that the official code of SR3 is yet to be released, and hence we resort to unofficial re-implementation, which we train with default configurations. Additionally, in the original work of SR3 [25], the authors propose consecutively applying $\times 4$ SR models to achieve $16 \times 16 \mapsto 64 \times 64 \mapsto 256 \times 256$ SR. In contrast, we report on a single $\times 16$ SR model which maps $16 \times 16 \mapsto 256 \times 256$ directly.

Figure 3. Stability of convergence depending on the choice of initialization. (a) Random initialization, large ε_0, (b) vanilla initialization, moderate ε_0, (c) NN initialization, small ε_0.

Inpainting. The score function used in the inpainting task is the same model that was used to solve SR tasks, since we use task-agnostic conditional diffusion model. The feed-forward network was adopted from Yu *et al.* [40]. We consider box-type inpainting with varying sizes: $96 \times 96, 128 \times 128, 160 \times 160$. The model was trained for 50k steps with default configurations. We compare with score-SDE [34], using the same trained score function.

MRI reconstruction. Experiments were performed across three different levels of acceleration factor, with gaussian 1D sampling pattern - $\times 2, \times 4, \times 6$, each with 10%, 8%, 6% of the phase encoding lines included for autocalibrating signal (ACS) region. We train a VE-SDE based on `ncsnpp`, proposed in [34], and demonstrated specifically for MR reconstruction in [6, 33]. For comparison with compressed sensing (CS) strategy, we use total-variation (TV) regularized reconstruction. For feed forward network, we train a standard U-Net, using similar settings from [6, 41]. We use the same trained score function for comparison with score-MRI [6].

[1] https://github.com/mseitzer/pytorch-fid

[2] https://github.com/Janspiry/Image-Super-Resolution-via-Iterative-Refinement

Figure 4. Results of super-resolution on AFHQ 256×256 data. First, second and third row denote ×4 SR, ×8 SR, and ×16 SR, respectively. (a) LR input, (b) Ground Truth, (c) ESRGAN [37], (d) SR3 [25] with 20 diffusion steps ($N = 20, \Delta t = 0.05$), (e) ILVR [5] with 20 diffusion steps ($N = 20, \Delta t = 0.05$), (f) proposed method with 20 diffusion steps ($N = 100, t_0 = 0.2$).

4.2. Super-resolution

Dependence on ε_0. We first demonstrate the dependency of stochastic contraction on the squared error term in Figure 3. For small squared difference, as in the case for many inverse problems, we see that the reverse diffusion stably converges to the same solution, even with small timestep t_0. In contrast, when random $\boldsymbol{x}_0$ is the starting point, ε_0 becomes large, and only with higher values of t_0 does the reverse SDE converge to a feasible solution.

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

Table 2. 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.

Dependence on t_0. In Table 2, we report on the FID scores by varying the t_0 values with a fixed discretization step $\Delta t = 1/1000$ in order to see which value is optimal for each degradation factor. Consistent with the theoretical findings, we see that as the corruption factor gets higher, and ε_0 gets larger, we typically need higher values of t_0 to achieve optimal results. Interesting enough, we observe that there *always* exist a value $t_0 \in [0, 1)$ where the FID score is lower (lower is better) than when using full reverse diffusion from $T = 1$.

Comparison study. The results of various super-resolution algorithms is compared in Fig. 4. We compare with SR3 [25] and ILVR [5], with setting the number of iterations for reconstruction same for ILVR, SR3, and the proposed method. We clearly see that SR3 and ILVR starting from pure Gaussian noise at $T = 1$ cannot generate satisfactory results with 20 iterations, whereas our method can estimate high-fidelity samples with details preserved even with only 20 iterations starting from $t_0 = 0.2$. Visualizing

the trend of FID score in Figure 5, we see that the quality of the image degrades as we use less and less number of iterations for the ILVR method, whereas the proposed method is able to keep the FID score at about the same level, or even *boost* the image quality, with less iterations.

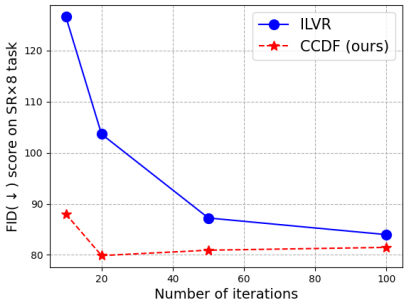

Figure 5. Comparison of FID score on $\times 8$ SR task. For ILVR, re-scheduling method of IDDPM [20] was used starting from $T = 1$. For CCDF, the step size for discretization is $\Delta t = 0.01$ so that the starting point for the reverse diffusion is $t_0 = \Delta t \times$ [number of iteration].

	SR factor	ESRGAN [37]	SR3* [25]	ILVR [5]	CCDF (ours)
FFHQ	$\times 4$	81.14	66.79	63.14	**60.90**
	$\times 8$	108.96	80.27	81.85	**75.76**
	$\times 16$	143.80	99.46	92.32	**88.39**
AFHQ	$\times 4$	24.52	20.68	18.70	**15.53**
	$\times 8$	51.84	30.23	34.85	**32.30**
	$\times 16$	98.22	60.76	**47.28**	48.77

Table 3. 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)

We also perform a comparison study where we set the total number of diffusion steps to $N = 1000$ starting from $T = 1$ for ILVR [5], and set t_0 to 0.1, 0.2, and 0.3 for each factor, thereby reducing the number of diffusion steps to 100, 200, and 300, respectively, by our method. In Table 3, we demonstrate by using the proposed method, we achieve results that are on par or even better. For qualitative analysis, see Supplementary Section E.

Incorporation of DDIM. As briefly discussed before, CCDF can be combined together with approaches that searches for the optimal (full) reverse diffusion path. In Fig. 6, we illustrate that we can reduce the number of iterations to as little as *5 steps*, and still maintain high image quality.

4.3. Inpainting

We illustrate the results of inpainting in Fig. 7. Consistent with what was observed in the SR task, the results in Figure 7 show that using full reverse diffusion with large discretization steps is inefficient, leading to unrealistic output. On the other hand, our method can reconstruct very

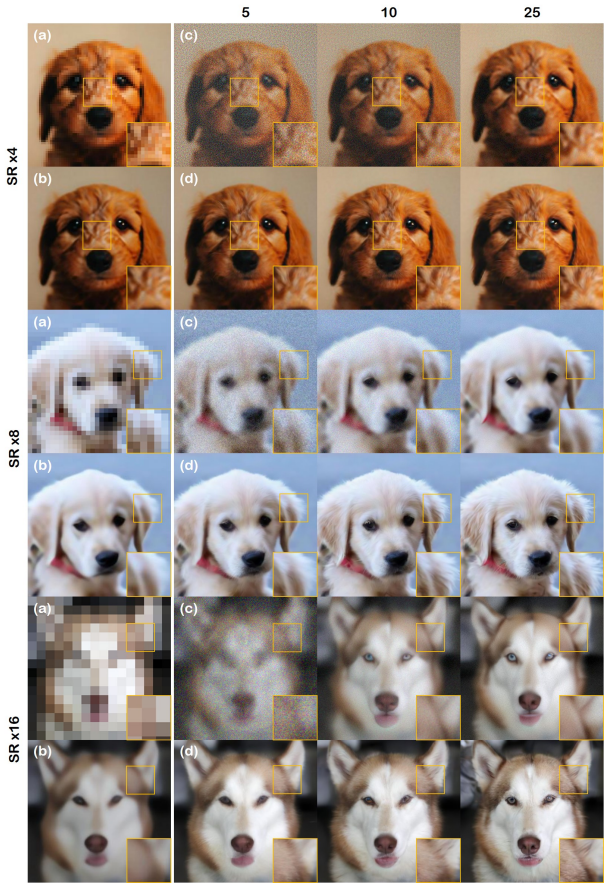

Figure 6. Results on SR task using CCDF with DDIM. (a) LR image, (b) initialization with ESRGAN, (c) ILVR + DDIM, (d) CCDF + DDIM. Numbers on top indicate the number of iterations. Proposed method uses $N = 50$, and $t_0 = 0.1, 0.2, 0.5$, respectively.

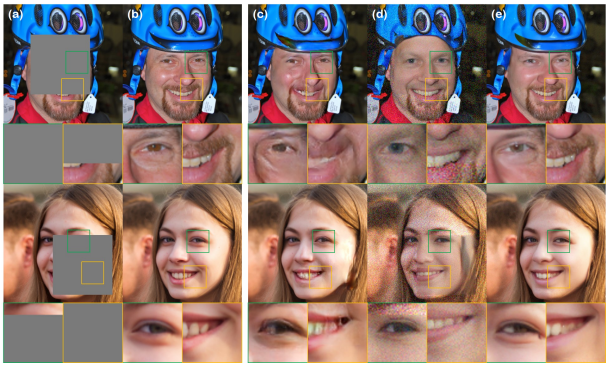

Figure 7. Results of inpainting on FFHQ 256×256 data. (a) Masked image, (b) Ground Truth, (c) SN-PatchGAN [40]. (d) score-SDE [34] using 20 steps from $T = 1$, (e) proposed method (CCDF) using 20 steps from $t_0 = 0.2$.

realistic images within this small budget.

method	masked	SN-PatchGAN [40]	Score-SDE [34] (1000)	CCDF (200)
Box 96	131.31	46.42	50.85	**45.99**
Box 128	145.81	52.63	64.51	**49.77**
Box 160	167.37	66.25	78.29	**57.99**

Table 4. Comparison of FID(↓) scores on FFHQ test set for inpainting task ($N = 1000, \Delta t = 0.001$). Number in parenthesis indicate the number of iterations used for generation. Numbers in boldface and underline indicate the best and the second best.

Comparison with prior arts by setting relatively large number of iterations is shown in Table 4. We observe that the proposed method outperforms both score-SDE with full reverse diffusion, and SN-PatchGAN, in terms of FID score. For detailed comparison and further experiments, see Supplementary Section E.

4.4. MRI reconstruction

We summarize and compare our results in Figure 8, and the quantitative metrics are presented in Table 5. In the task of MR reconstruction, we observe that we can push the t_0 value down to very small values: $t_0 = 0.02$, and still achieve remarkable results, even outperforming score-POCS which uses full reverse diffusion. When we compare the proposed method which uses 20 iterations vs. score-POCS with 20 iterations, we see that score-POCS cannot generate a feasible image, arriving at what looks like pure noise, as demonstrated in Figure 1. With other tasks, we could see that higher degradations typically require increased t_0 values. With CCDF, we do not see such trend, and observe that selecting low values of $t_0 \in [0.02, 0.1]$ stably gives good results. We emphasize that this is a huge leap towards practical usage of diffusion models in clinical settings, where fast reconstruction is crucial for real-time deployment.

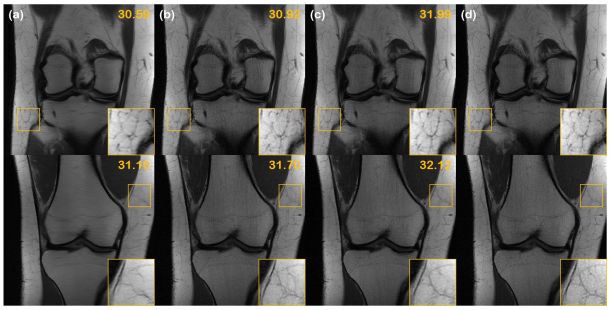

Figure 8. Results of the MR reconstruction task: (a) TV [3], (b) U-Net [41], (c) score-POCS [6] using 1000 steps starting from $T = 1$, (d) proposed method (CCDF) using 20 steps from $t_0 = 0.02$ (20 steps), (e) Reference image. Numbers in yellow correspond to PSNR values.

method	ZF	TV [3]	U-Net [41]	Score-POCS [6]	CCDF (20)
× 2	27.23	29.10	32.93	32.85	**33.41**
× 4	22.68	25.93	31.07	31.45	**32.51**
× 6	21.54	24.69	30.77	31.15	**31.30**

Table 5. PSNR(↑) on fastMRI test set for MRI reconstruction tasks. Gaussian 1D sampling masks were used. Number in parenthesis indicate the number of iterations used ($N = 1000, t_0 = 0.02$). Numbers in boldface indicate the best among the rows.

5. Discussion

We note that we are not the first to propose starting from forward-diffused data in the context of diffusion models. It was first introduced in SDEdit [19], but in a different context with distinct aim form ours. In SDEdit, forward diffusion was used up to $t_0 \in [0.3, 0.6]$, which a relatively higher value than those used in our work $t_0 \leq 0.2$, since the purpose was to *destroy* the signal so as to acquire high fidelity images from coarse strokes.

Our work differs from SDEdit in that we consider this procedure in a more rigorous framework and first reveal that starting from a better initialization for inverse problems significantly accelerate the reverse diffusion. This leads to a novel hybridization that has not been covered before: a simple incorporation of pre-trained feed-forward NNs can be very efficient at pushing t_0 to smaller limits, even as small as $t_0 = 0.02$ in the case of MRI reconstruction.

5.1. Limitations

We note that the choice of t_0 for acceleration varies by quite a margin across different tasks, and the degree of corruptions. Currently, there does not exist clear and concise rules for selecting such values as we do not have a knowledge of ε_0 a priori. Thus, one needs to rely mostly on trial-and-error, which could potentially reduce practicality. Building an adaptive method that can automatically search for the optimal t_0 values will be beneficial, and we leave this venue for possible direction of future research.

6. Conclusion

In this work, we proposed a method to accelerate conditional diffusion models, by studying the property of stochastic contraction. When solving inverse problems via conditional reverse diffusion, rather than starting at random Gaussian noise, we proposed to initialize the starting from forward-diffused data from a better initialization, such as one-step correction via NN. Using the stochastic contraction theory, we showed theoretically why taking the shortcut path is in fact optimal, and back our statement by showing diverse applications in which we both achieve acceleration along with increased stability and performance.

References

[1] Guillaume Alain and Yoshua Bengio. What regularized auto-encoders learn from the data-generating distribution. *The Journal of Machine Learning Research*, 15(1):3563–3593, 2014. 19

[2] Heinz H Bauschke, Patrick L Combettes, et al. *Convex analysis and monotone operator theory in Hilbert spaces*, volume 408. Springer, 2011. 4

[3] Kai Tobias Block, Martin Uecker, and Jens Frahm. Undersampled radial MRI with multiple coils. Iterative image reconstruction using a total variation constraint. *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 57(6):1086–1098, 2007. 8

[4] Nanxin Chen, Yu Zhang, Heiga Zen, Ron J Weiss, Mohammad Norouzi, and William Chan. WaveGrad: Estimating gradients for waveform generation. In *International Conference on Learning Representations*, 2021. 2

[5] Jooyoung Choi, Sungwon Kim, Yonghyun Jeong, Youngjune Gwon, and Sungroh Yoon. ILVR: Conditioning method for denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 1, 2, 6, 7, 15, 17, 22, 23

[6] Hyungjin Chung and Jong Chul Ye. Score-based diffusion models for accelerated mri. *arXiv preprint arXiv:2110.05243*, 2021. 1, 2, 5, 8, 16

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 18

[8] Prafulla Dhariwal and Alex Nichol. Diffusion models beat GANs on image synthesis. In *Advances in Neural Information Processing Systems*, 2021. 1, 3, 15

[9] Chong Fan, Chaoyun Wu, Grand Li, and Jun Ma. Projections onto convex sets super-resolution reconstruction based on point spread function estimation of low-resolution remote sensing images. *Sensors*, 17(2):362, 2017. 4

[10] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851, 2020. 1, 3, 15

[11] Hossein Hosseini, Neda Barzegar Marvasti, and Farrokh Marvasti. Image inpainting using sparsity of the transform domain. *arXiv preprint arXiv:1011.5458*, 2010. 4

[12] Chin-Wei Huang, Jae Hyun Lim, and Aaron Courville. A variational perspective on diffusion-based generative models and score matching. In *Advances in Neural Information Processing Systems*, 2021. 2, 3

[13] Ajil Jalal, Marius Arvinte, Giannis Daras, Eric Price, Alexandros G Dimakis, and Jonathan Tamir. Robust compressed sensing mri with deep generative priors. *Advances in Neural Information Processing Systems*, 34, 2021. 2, 4

[14] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR*, 2015. 15

[15] Diederik P Kingma, Tim Salimans, Ben Poole, and Jonathan Ho. Variational diffusion models. *Advances in Neural Information Processing Systems*, 34, 2021. 1, 2, 3

[16] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4681–4690, 2017. 5

[17] Haoying Li, Yifan Yang, Meng Chang, Huajun Feng, Zhihai Xu, Qi Li, and Yueting Chen. SRDiff: Single image super-resolution with diffusion probabilistic models. *arXiv preprint arXiv:2104.14951*, 2021. 2, 4

[18] Eric Luhman and Troy Luhman. Knowledge distillation in iterative generative models for improved sampling speed. *arXiv preprint arXiv:2101.02388*, 2021. 2

[19] Chenlin Meng, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Image synthesis and editing with stochastic differential equations. *CoRR*, abs/2108.01073, 2021. 2, 8

[20] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8162–8171. PMLR, 2021. 3, 5, 7, 15

[21] Sung Woo Park, Dong Wook Shu, and Junseok Kwon. Generative adversarial networks for markovian temporal dynamics: Stochastic continuous data generation. In *International Conference on Machine Learning*, pages 8413–8421. PMLR, 2021. 2

[22] Quang-Cuong Pham. Analysis of discrete and hybrid stochastic systems by nonlinear contraction theory. In *2008 10th International Conference on Control, Automation, Robotics and Vision*, pages 1054–1059. IEEE, 2008. 2, 4, 11, 13

[23] Quang-Cuong Pham, Nicolas Tabareau, and Jean-Jacques Slotine. A contraction theory approach to stochastic incremental stability. *IEEE Transactions on Automatic Control*, 54(4):816–820, 2009. 2, 4

[24] Zaccharie Ramzi, Benjamin Remy, Francois Lanusse, Jean-Luc Starck, and Philippe Ciuciu. Denoising score-matching for uncertainty quantification in inverse problems. In *NeurIPS 2020 Workshop on Deep Learning and Inverse Problems*, 2020. 4

[25] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *arXiv preprint arXiv:2104.07636*, 2021. 2, 4, 5, 6, 7, 22

[26] Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. In *International Conference on Learning Representations*, 2022. 2

[27] Alexei A Samsonov, Eugene G Kholmovski, Dennis L Parker, and Chris R Johnson. POCSENSE: POCS-based reconstruction for sensitivity encoded magnetic resonance imaging. *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 52(6):1397–1406, 2004. 4

[28] Hiroshi Sasaki, Chris G Willcocks, and Toby P Breckon. UNIT-DDPM: Unpaired image translation with denoising diffusion probabilistic models. *arXiv preprint arXiv:2104.05358*, 2021. 2

[29] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pages 2256–2265. PMLR, 2015. 1

[30] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *9th International Conference on Learning Representations, ICLR*, 2021. 2, 16

[31] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In *Advances in Neural Information Processing Systems*, volume 32, 2019. 1, 2, 3

[32] Yang Song and Stefano Ermon. Improved techniques for training score-based generative models. In *Advances in Neural Information Processing Systems*, volume 33, pages 12438–12448, 2020. 1, 3

[33] Yang Song, Liyue Shen, Lei Xing, and Stefano Ermon. Solving inverse problems in medical imaging with score-based generative models. In *International Conference on Learning Representations*, 2022. 2, 5

[34] Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *9th International Conference on Learning Representations, ICLR*, 2021. 1, 2, 3, 5, 7, 8, 12, 16, 18, 24

[35] Zhifei Tang, Mike Deng, Chuangbai Xiao, and Jing Yu. Projection onto convex sets super-resolution image reconstruction based on wavelet bi-cubic interpolation. In *Proceedings of 2011 International Conference on Electronic & Mechanical Engineering and Information Technology*, volume 1, pages 351–354. IEEE, 2011. 4

[36] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. 15

[37] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. Esrgan: Enhanced super-resolution generative adversarial networks. In *Proceedings of the European conference on computer vision (ECCV) workshops*, pages 0–0, 2018. 5, 6, 7, 17, 22, 23

[38] Daniel Watson, Jonathan Ho, Mohammad Norouzi, and William Chan. Learning to efficiently sample from diffusion probabilistic models. *arXiv preprint arXiv:2106.03802*, 2021. 2

[39] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Generative image inpainting with contextual attention. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5505–5514, 2018. 5

[40] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Free-form image inpainting with gated convolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4471–4480, 2019. 5, 7, 8, 18, 24

[41] Jure Zbontar, Florian Knoll, Anuroop Sriram, Tullie Murrell, Zhengnan Huang, Matthew J Muckley, Aaron Defazio, Ruben Stern, Patricia Johnson, Mary Bruno, et al. fastMRI: An open dataset and benchmarks for accelerated MRI. *arXiv preprint arXiv:1811.08839*, 2018. 5, 8