Zero-Shot Image Restoration Using Few-Step Guidance of Consistency Models (and Beyond)

Tomer Garber
Bar-Ilan University, Israel
tomergarber@gmail.com

Tom Tirer
Bar-Ilan University, Israel
tirer.tom@gmail.com

Abstract

In recent years, it has become popular to tackle image restoration tasks with a single pretrained diffusion model (DM) and data-fidelity guidance, instead of training a dedicated deep neural network per task. However, such "zero-shot" restoration schemes currently require many Neural Function Evaluations (NFEs) for performing well, which may be attributed to the many NFEs needed in the original generative functionality of the DMs. Recently, faster variants of DMs have been explored for image generation. These include Consistency Models (CMs), which can generate samples via a couple of NFEs. However, existing works that use guided CMs for restoration still require tens of NFEs or fine-tuning of the model per task that leads to performance drop if the assumptions during the fine-tuning are not accurate. In this paper, we propose a zero-shot restoration scheme that uses CMs and operates well with as little as 4 NFEs. It is based on a wise combination of several ingredients: better initialization, back-projection guidance, and above all a novel noise injection mechanism. We demonstrate the advantages of our approach for image super-resolution, deblurring and inpainting. Interestingly, we show that the usefulness of our noise injection technique goes beyond CMs: it can also mitigate the performance degradation of existing guided DM methods when reducing their NFE count. The source code is available at https://github.com/tirer-lab/CM4IR

1. Introduction

Image restoration problems appear in a wide range of applications, where the goal is to recover a high-quality image $\mathbf{x}^* \in \mathbb{R}^n$ from its degraded version $\mathbf{y} \in \mathbb{R}^m$, which can be incomplete, noisy, blurry, low-resolution, etc. In many popular tasks, the relation between $\mathbf{y}$ and $\mathbf{x}^*$ can be expressed using a linear observation model

$$\mathbf{y} = \mathbf{A}\mathbf{x}^* + \mathbf{e} \tag{1}$$

Figure 1. Super-resolution $\times 4$ with bicubic kernel and noise level of 0.05. From left to right and top to bottom: original, observation, DPS [6] (1000 NFEs), DiffPIR [40] (20 NFEs), DDRM [15] (20 NFEs) and our CM4IR (4 NFEs).

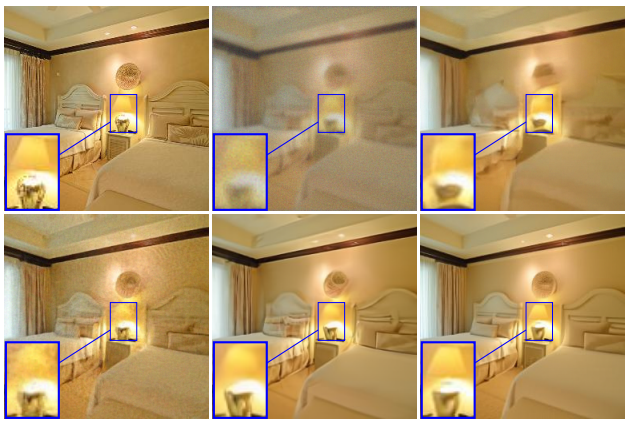

Figure 2. Deblurring with Gaussian kernel and noise level of 0.025. From left to right and top to bottom: original, observation, DPS [6] (1000 NFEs), DiffPIR [40] (20 NFEs), DDRM [15] (20 NFEs) and our CM4IR (4 NFEs).

where $\mathbf{A} \in \mathbb{R}^{m \times n}$ is a measurement operator with $m \leq n$, and $\mathbf{e} \sim \mathcal{N}(\mathbf{0}, \sigma_y^2 \mathbf{I}_m)$ is an additive white Gaussian noise. Examples include: denoising, inpainting, deblurring, super-resolution, computed tomography, magnetic resonance imaging, and more. Each of these tasks is associated with a different structure of $\mathbf{A}$. Importantly, image restoration problems are ill-posed: fitting the observation $\mathbf{y}$ does not ensure a successful recovery and thus prior knowledge on the nature of $\mathbf{x}^*$ needs to be utilized.

Since the breakthroughs in deep learning a decade ago [16], it has become common to address these problems by training a different deep neural network (DNN) for each *predefined* observation model in a supervised manner [7, 17]. Specifically, given a collection of high-quality images $\{\mathbf{x}_i^*\}$, training pairs $\{\mathbf{y}_i, \mathbf{x}_i^*\}$ are generated by mapping $\mathbf{x}_i^*$ to $\mathbf{y}_i$ using (1), and a DNN is trained to invert the map. However, these "task-specific" DNNs suffer from a huge performance drop when the observations at test-time mismatch (even slightly) the assumptions made in training [2, 21, 29, 32], which limits their applicability in many practical cases.

An increasingly popular alternative to "task specific" DNNs are "zero-shot" approaches, where a pretrained DNN imposes only the signal prior while agreement of the reconstruction with the observations (of potentially arbitrary model) is handled at test-time. Different choices of such DNNs include plain denoisers [35] as used in [28, 33, 36], generative adversarial networks [10] as used in [5, 13], and recently (unconditional) diffusion/score models (DMs) [12, 24] as used in [3, 6, 9, 15, 23, 25, 34, 40].

Despite their flexibility to the observation model, the main limitation of zero-shot restoration schemes that are based on DMs is that they require many Neural Function Evaluations (NFEs) of highly overparameterized models. The core reason for this is inherited from the original generative functionality of the DMs, which itself requires many NFEs. Indeed, the restoration schemes are based on equipping sampling schemes, consisting of many iterations of denoising and noise injection, with data-fidelity guidance. For a given DM, sophisticated guidance techniques can reduces the number of NFEs from 1000 NFEs [6] to a couple of dozens [15, 40], but to our knowledge not less than this. More details on the connections between types of data-fidelity guidance, their acceleration aspect, and the relation of guided DMs to proximal algorithms can be found in [9].

Attempts to reduce the number of NFEs required for image generation is an active research topic. Some works focus on modifying the sampling schemes ("reversed flows") of pretrained DM [14, 22]. A greater acceleration is achieved using new training strategies for the denoising models [1, 20, 27]. Among those, a prominent alternative to DMs are Consistency Models (CMs) [27], which can generate samples via a couple of NFEs. However, existing works that use guided CMs for restoration still require tens of NFEs [27] or fine-tuning of the model per task [39]. Note that the latter is not a zero-shot strategy and, as such, [39] is prone to suffer from significant performance drop if the assumptions during the fine-tuning (e.g., the noise level) are not accurate.

Contribution. In this paper, we propose a zero-shot restoration scheme that utilizes CMs and, with as little as 4 NFEs, outperforms alternative zero-shot methods (which use many more NFEs) on popular tasks such as image super-resolution, deblurring and inpainting. Our approach is based on a wise combination of several ingredients: better initialization, back-projection guidance, and above all a novel noise injection mechanism.

Our noise injection mechanism deviates from existing techniques in two aspects that allow us to reduce the number of iterations. First, we decouple the noise level of the denoising operation from the noise level of the injection, such that it takes into account noise that originates from the data-fidelity guidance and gives the denoiser more "freedom" to modify its input. Second, we split the amount of injected noise between the stochastic noise and an estimated noise (*anti*-correlated to the one in [22]) that boosts the sampling scheme. We formally motivate this split by: 1) showing that it preserves properties of marginal distribution of the samples when omitting the guidance (the aforementioned noise decoupling takes the guidance into account), and 2) identifying it as a "noisy variant" of Polyak acceleration [18].

Finally, we show that the usefulness of our noise injection technique *goes beyond CMs*: it can also mitigate the performance degradation of existing guided DM methods [15, 40] when the number of iterations (NFEs) that they use is drastically reduced.

2. Background and Related Work

2.1. DMs and CMs

Let $p_{\text{data}}(\mathbf{x})$ denote the data distribution. Diffusion/score-based generative models [12, 14, 24, 26] are based on reversing a user-designed "forward flow" where p_{data} is diffused till it reaches the tractable Gaussian distribution $\mathcal{N}(\mathbf{0}, \sigma_{\text{max}}^2 \mathbf{I})$. The forward flow can be expressed by the stochastic differential equation (SDE)

$$d\mathbf{x}_t = \mathbf{h}(\mathbf{x}_t, t)\mathrm{d}t + g(t)\mathrm{d}\mathbf{w}_t, \tag{2}$$

where $t \in [0, T]$ with constant $T > 0$, $\mathbf{h}$ and g are the drift and diffusion coefficients, respectively, and $\{\mathbf{w}_t\}$ denotes the standard Brownian motion. We denote the distribution of $\mathbf{x}_t$ by $p_t(\mathbf{x})$, so $p_0 = p_{\text{data}}$ and we aim for $p_T = \mathcal{N}(\mathbf{0}, \sigma_{\text{max}}^2 \mathbf{I})$.

As pointed out by [26], based on a result of [4], the for-

ward SDE is associated with a reversed SDE

$$d\mathbf{x}_t = \left[\mathbf{h}(\mathbf{x}_t, t) - g(t)^2 \nabla_{\mathbf{x}} \log p_t(\mathbf{x}_t)\right] \mathrm{d}t + g(t) \mathrm{d}\tilde{\mathbf{w}}_t, \tag{3}$$

where $\tilde{\mathbf{w}}$ is a standard Brownian motion when time flows backwards from T to 0, and $\mathrm{d}t$ is an infinitesimal negative timestep. Interestingly, [26] showed that the reversed SDE has an associated ordinary differential equation (ODE), dubbed Probability Flow (PF), whose solution trajectories sampled at t are distributed according to $p_t(\mathbf{x})$:

$$d\mathbf{x}_t = \left[\mathbf{h}(\mathbf{x}_t, t) - \frac{1}{2} g(t)^2 \nabla_{\mathbf{x}} \log p_t(\mathbf{x}_t)\right] \mathrm{d}t. \tag{4}$$

One can approximately obtain a sample of p_{data} from an initial $\mathbf{x}_T \sim p_T$ in a "reversed flow", provided that he has an accurate estimator of $\nabla_{\mathbf{x}} \log p_t(\cdot)$, known as the *score function* of p_t, and sufficiently good discretization and integration of the reversed SDE/ODE.

In DMs, a common choice is $\mathbf{h}(\mathbf{x}_t, t) = \mathbf{0}$, known as the VE-SDE formulation [14, 26]. In this case, by (2), $\mathbf{x}_t$ is a noisy version of an unscaled data sample $\mathbf{x}_0 \sim p_0$ with additive Gaussian noise of some level σ_t, depending on the choice of $g(t)$. For $\mathbf{x}_t = \mathbf{x}_0 + \sigma_t \boldsymbol{\epsilon}_t$ with $\boldsymbol{\epsilon}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, we have that $-\sigma_t^2 \nabla_{\mathbf{x}} \log p_t(\mathbf{x}_t) = \mathbf{x}_t - \mathbb{E}[\mathbf{x}_0|\mathbf{x}_t] = \sigma_t \mathbb{E}[\boldsymbol{\epsilon}_t|\mathbf{x}_t]$, where the first equality is Tweedie's formula [8] and the second equality follows from $\mathbb{E}[\mathbf{x}_0|\mathbf{x}_t] = \mathbb{E}[\mathbf{x}_t - \sigma_t \boldsymbol{\epsilon}_t|\mathbf{x}_t] = \mathbf{x}_t - \sigma_t \mathbb{E}[\boldsymbol{\epsilon}_t|\mathbf{x}_t]$. Therefore, conceptually, there is equivalence between learning: 1) the score function [24, 26], 2) the MMSE denoiser [14], or 3) the MMSE of the noise [12]. Similar equivalence, up to scaling factors, holds in the VP-SDE formulation, where $\mathbf{h}(\mathbf{x}_t, t) = -\frac{1}{2}\beta(t)\mathbf{x}_t$ and $g(t) = \sqrt{\beta(t)}$ [12, 26].

A denoiser $f_\theta(\mathbf{x}, \sigma)$, where $\mathbf{x}$ is the signal and σ is the noise level, can be parameterized in different ways. For example, [14] parameterized it as

$$f_\theta(\mathbf{x}, \sigma) = c_{\text{skip}}(\sigma)\mathbf{x} + c_{\text{out}}(\sigma) F_\theta\left(c_{\text{in}}(\sigma)\mathbf{x}, c_{\text{noise}}(\sigma)\sigma\right), \tag{5}$$

where F_θ is the DNN and $c_{\text{skip}}(\cdot), c_{\text{out}}(\cdot), c_{\text{in}}(\cdot), c_{\text{noise}}(\cdot)$ are coefficient functions that are set manually. Their denoiser is trained for the range of noise level σ that is required for the reversed flow by minimization of

$$\mathbb{E}_{\sigma, \mathbf{x}_0 \sim p_{\text{data}}, \mathbf{x} \sim \mathcal{N}(\mathbf{x}_0, \sigma^2 \mathbf{I})} \left[\lambda(\sigma) d\left(f_\theta(\mathbf{x}, \sigma), \mathbf{x}_0\right)\right] \tag{6}$$

where $\lambda(\sigma)$ is a weighting function the loss and the distance metric is $d(\mathbf{x}, \tilde{\mathbf{x}}) = \|\mathbf{x} - \tilde{\mathbf{x}}\|_2^2$, aligned with MMSE estimation.

After training, the sampling schemes are based on numerical solvers (discretization and integration) of the reversed SDE (3), which boil down to random Gaussian initialization of $\mathbf{x}_T$ and iterations of denoising and noise injection with decreasing noise levels. Many works also use the choice $g(t) = \sqrt{2t}$, which leads to a simple $\sigma_t = t$ [14, 26, 27]. Then, for a sequence of time points $T = \tau_N > \tau_{N-1} > \ldots > \tau_2 > \tau_1 = \epsilon$ (discretization of $[\epsilon, T]$ with $\epsilon \approx 0_+$ and typically $T = 1$) and initialization $\mathbf{x}_{\tau_N} \sim \mathcal{N}(\mathbf{0}, \tau_N^2 \mathbf{I})$, a typical sampling scheme with $n = N, N-1, ..., 1$ reads as

$$\mathbf{x}_{0|\tau_n} = f_\theta(\mathbf{x}_{\tau_n}, \tau_n) \tag{7}$$

$$\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \tag{8}$$

$$\mathbf{x}_{\tau_{n-1}} = \mathbf{x}_{0|\tau_n} + \tau_{n-1}\mathbf{z} \tag{9}$$

Attempts to reduce the overall number of NFEs in the sampling scheme include improving the estimated denoised signal using two NFEs per iteration [14] (i.e., second-order integrator) or replacing the noise injection in (9) with the DDIM procedure [22]:

$$\sqrt{1 - \eta^2}\tau_{n-1}\hat{\mathbf{z}} + \eta\tau_{n-1}\mathbf{z}, \tag{10}$$

where $\eta \in [0, 1]$ is a hyperparameter that trades between the stochastic noise $\mathbf{z}$ and the estimated noise

$$\hat{\mathbf{z}} = (\mathbf{x}_{\tau_n} - f_\theta(\mathbf{x}_{\tau_n}, \tau_n))/\tau_n. \tag{11}$$

Note (10) slightly differs from the presentation in [22], since they use VP-SDE formulation and a noise estimator model (i.e., $\mathbf{z}_\theta(\mathbf{x}_{\tau_n}, \tau_n)$ rather than $\hat{\mathbf{z}}$ and $f_\theta(\mathbf{x}_{\tau_n}, \tau_n)$).

However, despite these efforts, with a denoiser (or noise estimator) optimized as in (6), even techniques that use more sophisticated updates require not less than three dozens NFEs for high-quality image generation [14].

To achieve a significant decrease in the number of NFEs, recent works have proposed different training strategies for the denoising models [1, 20, 27]. Among them, Consistency Models (CMs) [27] has received much attention.

CMs attempt to train a denoiser f_θ in a way that better promotes that it outputs the same clean image $\mathbf{x}_0$ for different points along path of the PF ODE (4) from $\mathbf{x}_T$ to $\mathbf{x}_0$. To this end, the most performant training technique in [27] is based on distilling a pre-trained DM Φ to obtain by deterministic numerical integration $\hat{\mathbf{x}}^\Phi_{t_n}$, an approximation of the point $\mathbf{x}_{t_n}$ on the PF path from $\mathbf{x}_T$ to $\mathbf{x}_0$, given $\mathbf{x}_0$. For a predetermined sequence for training $T = t_N > t_{N-1} > \ldots > t_2 > t_1 = \epsilon$, which may differ from the one used for sampling, the common training loss (6) (with $\sigma = t$) is replaced with

$$\mathbb{E}_{n \sim U[1, N-1], \mathbf{x}_0 \sim p_{\text{data}}, \mathbf{x}_{t_{n+1}} \sim \mathcal{N}(\mathbf{x}_0, t_{n+1}^2 \mathbf{I})} \left[\lambda(t) d\left(f_\theta(\mathbf{x}_{t_{n+1}}, t_{n+1}), f_{\theta^-}(\hat{\mathbf{x}}^\Phi_{t_n}, t_n)\right)\right], \tag{12}$$

where f_{θ^-} is a "target" copy of the denoiser whose parameters θ^- are exponential moving average of θ with $\mathrm{stopgrad}$. To avoid degenerate constant output of f_θ, a specific parameterization (5) is used: $c_{\text{skip}}(\epsilon) = 1$ and $c_{\text{out}}(\epsilon) = 0$, which

ensures that $f_\theta(\mathbf{x}, \epsilon) = \mathbf{x}$. We refer the reader to [27] for additional details on the architecture, coefficients functions and the training procedure of CMs. We note that [27] proposes also an approach for training the denoiser without distillation, but the aforementioned approach performs better, so we choose to consider CMs trained in this way.

Remarkably, CMs have been shown to generate quite good images already with 1 NFE, i.e., simply by $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, T^2\mathbf{I})$ and $\mathbf{x}_{0|T} = f_\theta(\mathbf{x}_T, T)$. The results are improved when using the iterative sampling (7)-(9) even with very small N and well chosen $\{\tau_n\}$.

2.2. Restoration via Guidance of DMs

Pretrained DMs have been shown to be a powerful signal prior for image restoration [3, 6, 9, 15, 23, 25, 34, 40]. In order that the sampling scheme will produce an image that is not only perceptually pleasing but also agrees with the measurements $\mathbf{y}$ and the observation model (1), it is required to equip the iterations with some data-fidelity guidance. This guidance is typically based on the gradient of a data-fidelity term $\ell(\mathbf{x}; \mathbf{y})$. In other words, the update (9) (oftentimes with the DDIM noise injection (10) [9, 15, 34, 40]) is modified into

$$\mathbf{x}_{\tau_{n-1}} = \mathbf{x}_{0|\tau_n} - \mu_n \nabla_{\mathbf{x}} \ell(\mathbf{x}_{0|\tau_n}; \mathbf{y}) + \tau_{n-1}\mathbf{z}, \tag{13}$$

where μ_n is the guidance scaling factor.

Least squares (LS) guidance is a typical choice

$$\ell_{\text{LS}}(\mathbf{x}; \mathbf{y}) = \frac{1}{2}\|\mathbf{A}\mathbf{x} - \mathbf{y}\|_2^2, \tag{14}$$

$$\nabla_{\mathbf{x}} \ell_{\text{LS}}(\mathbf{x}; \mathbf{y}) = \mathbf{A}^T(\mathbf{A}\mathbf{x} - \mathbf{y}), \tag{15}$$

which is used, e.g., in [6]. Yet, acceleration (less NFEs) is obtained with the back-projection (BP) guidance [28, 30]

$$\ell_{\text{BP}}(\mathbf{x}; \mathbf{y}) = \frac{1}{2}\|(\mathbf{A}\mathbf{A}^T)^{-1/2}(\mathbf{A}\mathbf{x} - \mathbf{y})\|_2^2, \tag{16}$$

$$\nabla_{\mathbf{x}} \ell_{\text{BP}}(\mathbf{x}; \mathbf{y}) = \mathbf{A}^\dagger(\mathbf{A}\mathbf{x} - \mathbf{y}), \tag{17}$$

where $\mathbf{A}^\dagger = \mathbf{A}^T(\mathbf{A}\mathbf{A}^T)^{-1}$ is the pseudoinverse of $\mathbf{A}$, which can be regularized and oftentimes has efficient implementation (generally by conjugate gradients [11], and in super-resolution and deblurring even by FFT [9]). See [31] for theoretical analysis of the acceleration property. This guidance has been used recently in [9, 23, 25, 34], sometimes under different names (e.g., "pseudoinverse guidance"). When the SVD of $\mathbf{A}$ can be computed efficiently, a guidance that resembles BP but can mitigate noise amplification per singular component has been used in the DDRM method [15]. Importantly, existing zero-shot restoration techniques based on guided DMs still require at least a couple of dozens NFEs.

CMs pave the way to zero-shot restoration techniques with less NFEs. However, existing works that use guided CMs [27, 39] still have limitations. Specifically, the guided scheme in the original CM paper [27] requires 40 NFEs. In the very recent CoSIGN paper [39], the authors apply supervised fine-tuning to the model per task, which allows them to restore perceptually pleasing images with a couple of iterations via BP-guidance. Yet, each NFE in [39] is more computationally heavy as a ControlNet [37] is integrated to the original pretrained CM. Furthermore, and more crucially, this is not a zero-shot strategy and, as such, leads to performance drop if the assumptions during the fine-tuning (e.g., the noise level) mismatch the situation at test-time.

3. The Proposed Method

In this paper, we propose a few-step zero-shot restoration technique that utilizes pretrained CMs. Our approach combines several ingredients: initialization, BP guidance, and above all a novel noise injection mechanism, each is described below.

Initialization. Most of the DM based restoration techniques initialize $\mathbf{x}_{\tau_N}$ with pure noise $\mathbf{x}_{\tau_N} \sim \mathcal{N}(\mathbf{0}, T^2\mathbf{I})$, or equivalently

$$\mathbf{x}_{\tau_N} = \mathbf{x}_{\text{init}} + \tau_N \mathbf{z} \tag{18}$$

with $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, $\mathbf{x}_{\text{init}} = \mathbf{0}$ and $\tau_N = T$. This ignores the fact that the observations vector $\mathbf{y}$ contains information on the specific $\mathbf{x}^*$ that we wish to restore. Therefore, in general, we propose to set $\mathbf{x}_{\text{init}} = \mathbf{A}^\dagger\mathbf{y}$. For example, if $\mathbf{A}$ is bicubic downsampling (in super-resolution task), then $\mathbf{A}^\dagger$ is bicubic upsampling. For inpainting, though, $\mathbf{A}$ is subsampled rows of the identity matrix and $\mathbf{A}^\dagger = \mathbf{A}^T$ just fills the missing pixels with zeros, so we propose to use a better choice: median initialization as in [28]. Our initialization resembles to the common practice in many methods prior to DMs [28, 33, 36]. The difference from these works, though, is that we still add noise of level τ_N to $\mathbf{x}_{\text{init}}$. To avoid masking $\mathbf{x}_{\text{init}}$ at initialization, the noise level τ_N will be smaller than T.

Guidance. Following the acceleration property of back-projections [28, 30, 31], we choose to use the BP guidance (17). Note that at high noise levels in $\mathbf{y}$ this may require using decaying step sizes $\{\mu_n\}$.

Noise injection. We propose a novel noise injection mechanism that deviates from existing techniques in two aspects.

First, we claim that in restoration tasks, the noise level of the denoising operation and the noise level of the injection do not need to be the same. This is due to: 1) If $\mathbf{y}$ contains noise, $\sigma_y > 0$, then the guidance $\nabla_{\mathbf{x}} \ell(\mathbf{x}_{0|\tau_n}; \mathbf{y})$ adds to the sample $\mathbf{x}_{\tau_{n-1}}$ a noise component that originates from $\mathbf{y}$; and 2) Regardless of σ_y, at early iterations the estimated signal can differ significantly from the underlying signal $\mathbf{x}^*$, and thus by increasing the denoiser's noise level above the

value used for the injection noise, we give the denoiser more "freedom" in modifying its input, which in return allows us to reduce the number of iterations. Therefore, for noise injection of level τ_n we propose to apply the CM f_θ with noise level $(1+\delta)\tau_n$, where $\delta \geq 0$ is a hyperparameter.

Second, to further accelerate the restoration with CMs, ideally we would like to push the sample $\mathbf{x}_{\tau_n}$ towards $\mathbf{x}_0$, i.e., in the direction $\mathbf{x}_0 - \mathbf{x}_{\tau_n}$, beyond what we get from one denoising operation of CM. However, we do not have $\mathbf{x}_0$ but rather the denoiser's output $\mathbf{x}_{0|\tau_n}$. This motivates us to define:

$$\hat{\mathbf{z}}^- := (\mathbf{x}_{0|\tau_n} - \mathbf{x}_{\tau_n})/\tau_n. \tag{19}$$

For $\mathbf{x}_t = \mathbf{x}_0 + t\mathbf{z}$ with $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, we have that $\mathbb{E}[\mathbf{z}|\mathbf{x}_t] = \mathbb{E}[(\mathbf{x}_t - \mathbf{x}_0)/t|\mathbf{x}_t] = (\mathbf{x}_t - \mathbb{E}[\mathbf{x}_0|\mathbf{x}_t])/t$. Considering $\mathbf{x}_{0|t}$ as a replacement of $\mathbb{E}[\mathbf{x}_0|\mathbf{x}_t]$, we can understand $\hat{\mathbf{z}}^-$ as the *negative* estimate of the noise. Note that the sign does not affect the similarity of $\hat{\mathbf{z}}^-$ to Gaussian noise. Thus, we propose to split the noise injection between the stochastic noise $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and the estimated $\hat{\mathbf{z}}^-$. The proposed noise injection for generating $\mathbf{x}_{\tau_{n-1}}$ reads as

$$\sqrt{1-\eta^2}\tau_{n-1}\hat{\mathbf{z}}^- + \eta\tau_{n-1}\mathbf{z} \tag{20}$$

where $\eta \in [0, 1]$ is a hyperparameter that trades between the two types of noise.

The full algorithm. The proposed technique, CM for Image Restoration (CM4IR), is summarized in Algorithm 1. In Section 4 we will show its performance with a choice of $N = 4$ values of $\{\tau_n\}$, i.e., with 4 NFEs.

Relation of (20) to DDIM. As shown in (10), recall that the DDIM scheme [22] accelerates a baseline scheme [12] by noise injection similar to (20) but with $\hat{\mathbf{z}} = -\hat{\mathbf{z}}^-$ instead of $\hat{\mathbf{z}}^-$ (i.e., anti-correlation). In [22], $\hat{\mathbf{z}}$ is motivated as being a "direction pointing to the current $\mathbf{x}_{\tau_n}$". We claim that for few-step restoration, attracting the update back to the current $\mathbf{x}_{\tau_n}$, by using $\hat{\mathbf{z}}$ instead of $\hat{\mathbf{z}}^-$ is not beneficial. This will be demonstrated empirically in Section 4.1. All the more so, in Section 4.3 we will show that this simple modification mitigates the performance drop of guided DM methods [15, 40], which are based on DDIM scheme, when drastically reducing the number of the NFEs.

Furthermore, in [22], a major motivation for the procedure (10) (with $\eta < 1$) is that under their assumption on the distribution $q_\eta(\mathbf{x}_{\tau_1:\tau_N}|\mathbf{x}_0)$, for any $n > 1$ the distribution[1]

$$q_\eta(\mathbf{x}_{\tau_{n-1}}|\mathbf{x}_{\tau_n}, \mathbf{x}_0) = \mathcal{N}\left(\mathbf{x}_0 + \sqrt{1-\eta^2}\tau_{n-1}\frac{\mathbf{x}_{\tau_n} - \mathbf{x}_0}{\tau_n}, \eta^2\tau_{n-1}^2\mathbf{I}\right) \tag{21}$$

from which their reversed flow is derived (by replacing $\mathbf{x}_0$ with its current estimate) possesses the "desired" marginal $q_\eta(\mathbf{x}_{\tau_{n-1}}|\mathbf{x}_0) = \mathcal{N}(\mathbf{x}_0, \tau_{n-1}^2\mathbf{I})$, which matches pretrained DMs (no dependency on η). This property is preserved in our case, as stated below and proved in the supplementary material.

[1] We modified the notations of [22] from VP-SDE to VE-SDE.

Algorithm 1 CM for Image Restoration (CM4IR)

Require: $f_\theta(\cdot, t)$ (CM denoiser), N, $\{\tau_n\}$, $\{\mu_n\}$, δ, η, $\mathbf{A}$, $\mathbf{y}$.

1: Initialize $\mathbf{x}_{\tau_N} \sim \mathcal{N}(\mathbf{A}^\dagger\mathbf{y}, \tau_N^2\mathbf{I}_n)$
2: **for** n from N to 1 **do**
3: $\quad\mathbf{x}_{0|\tau_n} = f_\theta(\mathbf{x}_{\tau_n}, (1+\delta)\tau_n)$
4: $\quad\mathbf{g}_{\text{BP}} = \mathbf{A}^\dagger(\mathbf{A}\mathbf{x}_{0|t} - \mathbf{y})$
5: $\quad\hat{\mathbf{z}}^- = (\mathbf{x}_{0|\tau_n} - \mathbf{x}_{\tau_n})/\tau_n$
6: $\quad\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_n)$
7: $\quad\mathbf{x}_{\tau_{n-1}} = \mathbf{x}_{0|\tau_n} - \mu_n\mathbf{g}_{\text{BP}} + \sqrt{1-\eta^2}\tau_{n-1}\hat{\mathbf{z}}^- + \eta\tau_{n-1}\mathbf{z}$
8: **end for**
9: **return** $\mathbf{x}_{0|\tau_1}$

Proposition 3.1. *Under the assumption of [22] on $q_\eta(\mathbf{x}_{\tau_1:\tau_N}|\mathbf{x}_0)$, we have $q_\eta(\mathbf{x}_{\tau_{n-1}}|\mathbf{x}_0) = \mathcal{N}(\mathbf{x}_0, \tau_{n-1}^2\mathbf{I})$ also if we replace $(\mathbf{x}_{\tau_n} - \mathbf{x}_0)/\tau_n$ with $(\mathbf{x}_0 - \mathbf{x}_{\tau_n})/\tau_n$ in (21).*

Relation to Polyak acceleration.

In his seminal work [18], Polyak suggested to accelerate the gradient descent optimization of an objective, e.g., $\ell(\mathbf{x})$, by the iterates (with a decreasing sequence of indices $\tau_N, \tau_{N-1}, ..., \tau_1$):

$$\mathbf{x}_{\tau_{n-1}} = \mathbf{x}_{\tau_n} - \mu\nabla\ell(\mathbf{x}_{\tau_n}) + \beta(\mathbf{x}_{\tau_n} - \mathbf{x}_{\tau_{n+1}}) \tag{22}$$

where μ and β are step-size and momentum hyperparameters, respectively. With suitable choice of hyperparameters, faster convergence than plain gradient descent can be proved for quadratic objective functions [18, 19]. Empirically, the approach can be beneficial in other settings.

When using a prior, e.g., a denoiser $f_\theta(\cdot, t)$, a projected/proximal version of (22) reads as:

$$\mathbf{x}_{0|\tau_n} = f_\theta(\mathbf{x}_{\tau_n}, \tau_n) \tag{23}$$

$$\mathbf{x}_{\tau_{n-1}} = \mathbf{x}_{0|\tau_n} - \mu\nabla\ell(\mathbf{x}_{0|\tau_n}) + \beta(\mathbf{x}_{0|\tau_n} - \mathbf{x}_{0|\tau_{n+1}}) \tag{24}$$

Comparing (24) with Algorithm 1 (line 7), we see that apart from the stochastic noise $\mathbf{z}$, the main difference is that our $\hat{\mathbf{z}}^- \propto \mathbf{x}_{0|\tau_n} - \mathbf{x}_{\tau_n}$ while the direction of Polyak acceleration is $\mathbf{v} := \mathbf{x}_{0|\tau_n} - \mathbf{x}_{0|\tau_{n+1}}$. Notice that $\mathbf{x}_{\tau_n}$ can be understood as a noisy version of $\mathbf{x}_{0|\tau_{n+1}}$ — and thus $\hat{\mathbf{z}}^-$ is a noisy version of $\mathbf{v}$. We claim that such a noise component is essential for our restoration scheme. Indeed, with properly scaled noise, the input to the denoiser in the next iteration better matches the data that the denoiser has been trained on. This will be demonstrated empirically in Section 4.1 by replacing $\hat{\mathbf{z}}^-$ with $\beta\mathbf{v}$ in Algorithm 1 (line 7) and tuning β in addition to other hyperparameters.

Table 1. Ablation study on super-resolution with 4 NFEs. PSNR [dB] (↑) and LPIPS (↓) results on LSUN Bedroom validation set.

Task \ Method	Alg.1 with δ=0 and η=1	Alg.1 with δ=0 and $\hat{\mathbf{z}}$ i/o $\hat{\mathbf{z}}^-$	Alg.1 with δ=0 and $\beta\mathbf{v}$ (Polyak) i/o $\hat{\mathbf{z}}^-$	Alg.1 with δ=0	CM4IR
SRx4 σ_y=0.025	24.49 / 0.349	24.64 / 0.348	23.37 / 0.367	25.94 / 0.298	**26.14 / 0.295**
SRx4 σ_y=0.05	23.37 / 0.361	20.95 / 0.606	22.30 / 0.434	25.51 / **0.320**	**25.60 / 0.320**

Table 2. Super-resolution, deblurring and inpainting. PSNR [dB] (↑) and LPIPS (↓) results on LSUN Bedroom validation set.

Task \ Method	CM (40 NFEs)	CoSIGN (task spec.)	DDRM (20 NFEs)	DiffPIR (20 NFEs)	CM4IR (Ours, 4 NFEs)
SRx4 σ_y=0.025	24.66 / 0.344	26.10 / **0.205**	25.67 / 0.316	25.09 / 0.374	**26.14** / 0.295
SRx4 σ_y=0.05	23.62 / 0.449	20.35 / 0.569	25.08 / 0.354	23.83 / 0.457	**25.60 / 0.320**
Gauss. Deblurring σ_y=0.025	26.07 / 0.339	19.74 / 0.342	28.94 / 0.221	27.48 / 0.319	**29.00 / 0.213**
Gauss. Deblurring σ_y=0.05	24.18 / 0.453	19.08 / 0.543	27.35 / 0.280	26.14 / 0.363	**27.42 / 0.258**
Inpaint. (80%) σ_y=0	22.39 / 0.366	23.16 / 0.397	19.40 / 0.545	22.78 / 0.464	**25.43 / 0.284**
Inpaint. (80%) σ_y=0.025	22.17 / 0.417	23.22 / 0.368	19.16 / 0.548	22.65 / 0.477	**25.34 / 0.295**
Inpaint. (80%) σ_y=0.05	21.56 / 0.476	23.22 / 0.442	19.09 / 0.560	22.38 / 0.496	**25.28 / 0.328**

Table 3. Super-resolution, deblurring and inpainting. PSNR [dB] (↑) and LPIPS (↓) results on LSUN Cat validation set.

Task \ Method	CM (40 NFEs)	CoSIGN (task spec.)	DDRM (20 NFEs)	DiffPIR (20 NFEs)	CM4IR (Ours, 4 NFEs)
SRx4 σ_y=0.025	25.63 / 0.366	N/A	26.93 / 0.329	26.70 / 0.349	**27.18 / 0.328**
SRx4 σ_y=0.05	24.03 / 0.459	N/A	26.05 / 0.371	25.45 / 0.399	**26.53 / 0.349**
Gauss. Deblurring σ_y=0.025	26.69 / 0.346	N/A	**29.84** / 0.258	27.93 / 0.330	29.76 / **0.247**
Gauss. Deblurring σ_y=0.05	24.54 / 0.453	N/A	**28.33** / 0.316	26.64 / 0.370	28.06 / **0.299**
Inpaint. (80%) σ_y=0.025	21.89 / 0.478	N/A	18.51 / 0.648	22.78 / 0.498	**25.89 / 0.364**
Inpaint. (80%) σ_y=0.05	21.07 / 0.523	N/A	18.48 / 0.649	22.49 / 0.514	**25.34 / 0.423**

4. Experiments

In this section, we examine the performance of CM4IR. We apply it with $N = 4$, so per task there are 4 time points values $\{\tau_n\}_{n=1}^N$ (noise levels) that can be tuned. Note that other works that use CMs (for either generation or restoration) [27, 39] require careful tuning of the time points. Yet, to reduce the tuning effort, here we use only two hyperparameters for this goal: $\gamma > 0$ and small $\overline{\alpha}_N > 0$, and set $\overline{\alpha}_{n-1} = \overline{\alpha}_n(1+\gamma)$ and $\tau_n = \sqrt{1-\overline{\alpha}_n}$. As for the other hyperparameters, we generally use $\mu_n = 1, \eta = 0.1$.

Datasets. We consider the datasets LSUN bedroom and LSUN cat (256×256), for which there exits both well pretrained CMs [27] and DMs (specifically DDPMs) [12]. For the test images, we consider 300 validation images for each dataset (as done in [39]). In the supplementary material we also report results for ImageNet 64×64.

Tasks. We consider image super-resolution (SR), deblurring and inpainting tasks, which have been examined also in previous works, and for each of them we consider noise levels of 0.025 and 0.05. Specifically, we consider SR with bicubic downsampling of factor 4, and deblurring with Gaussian kernel as used, e.g., in [9, 15]. For the inpainting task, we consider the case of random 80% missing pixels and use median init for all the examined methods (we verified that this is beneficial for all of them). In the supplementary material we also present results for removal of superimposed text. For each competing method we use hyperparameter settings as suggested by its authors. The hyperparameter settings for our CM4IR are stated in the supplementary material. We evaluate the performance of methods by the PSNR and (VGG-)LPIPS [38] metrics.

4.1. Ablation study

We start by conducting an ablation study of CM4IR on the super-resolution tasks, where we disable parts of Algorithm 1. Specifically, we examine what happens if: 1) we do not utilize our noise injection strategy ($\delta = 0, \eta = 1$); 2) we use $\hat{\mathbf{z}} := -\hat{\mathbf{z}}^-$ (like DDIM [22]) instead of $\hat{\mathbf{z}}^-$; and 3) we use $\beta\mathbf{v}$ (i.e., plain Polyak acceleration with tunable β) instead of $\hat{\mathbf{z}}^-$. *Importantly:* in each case we tune the hyperparameters, and in particular the time points (via γ and $\overline{\alpha}_N$), separately for best performance.

The PSNR and LPIPS results are presented in Table 1. We see the benefit from having the different ingredients in CM4IR. In particular, using $\hat{\mathbf{z}}$ instead of $\hat{\mathbf{z}}^-$ does not improve much upon the $\eta = 1$ case. Using $\beta\mathbf{v}$ instead of $\hat{\mathbf{z}}^-$ even deteriorates the results compared to the $\eta = 1$ case. The advantage of using $\hat{\mathbf{z}}^-$ over $\hat{\mathbf{z}}$ implies that when having only 4 NFEs, one should not promote closeness of $\mathbf{x}_{\tau_{n-1}}$ to $\mathbf{x}_{\tau_n}$. Combining it with $\delta > 0$ improves the PSNR (higher gains have been observed for inpainting, as shown in the

Figure 3. SRx4 with noise level 0.05. From left to right: original, upsampled observation, DiffPIR (20 NFEs), DDRM (20 NFEs), CM (40 NFEs), CoSIGN (task specific) and our CM4IR (4 NFEs).

Figure 4. Inpainting (80% missing pixels) with noise level 0.05. From left to right: original, observation, DiffPIR (20 NFEs), CM (40 NFEs), CoSIGN (task specific) and our CM4IR (4 NFEs).

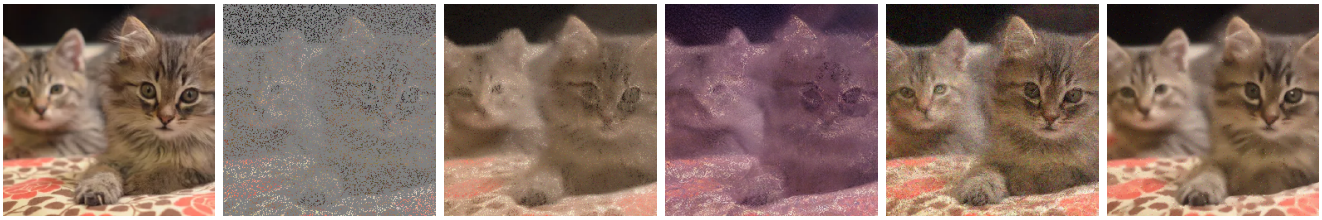

Figure 5. Inpainting (80% missing pixels) with noise level 0.05. From left to right: original, observation, DiffPIR (20 NFEs), DDRM (20 NFEs), CM (40 NFEs) and our CM4IR (4 NFEs).

supplementary material).

4.2. Comparison with other methods

We turn to compare CM4IR with other methods. We consider the guided CM scheme proposed in [27], which is based on a predetermined sequence $\{\tau_n\}$ of size $N = 40$ (i.e., 40 NFEs), and thus we name it CM(40). The guidance in this scheme resembles BP. The major difference of our CM4IR from CM(40) is the noise injection, which allows obtaining good results with a sequence $\{\tau_n\}$ only of size $N = 4$. We also present the results of CoSIGN [39], which is based on supervised fine-tuning to the model per task. For fair comparison, all the methods are applied with the same CMs.

We do not report results for CoSIGN on LSUN cat test set, because its additional ControlNet module is trained only on LSUN bedroom test set per task, and, unsurprisingly, it completely fails (less than 10dB PSNR) on LSUN cat even on the tasks considered in [39]. Another important remark on CoSIGN is that while the tasks in [39] are stated to include noise level $\sigma_y = 0.05$, when inspecting their code we saw that they double the range of the clean signal (to $[-1, 1]$) but did not double the noise level as done in the literature [9, 15, 40]. Therefore, this task-specific method is actually trained for $\sigma_y = 0.025$.

We compare our CM4IR also with zero-shot guided DM approaches: DDRM [15] and DiffPIR [40], which require 20 NFEs. We apply these methods with the same DMs downloaded from https://huggingface.co/google.

The results for the LSUN bedrom test set and the LSUN cat test set are presented in Table 2 and Table 3, respectively. We present qualitative results in Figures 1 - 5. More results appear in the supplementary material.

Examining the results, we start with comparing our CM4IR with CM(40) and CoSIGN. Clearly, CM4IR outperforms CM(40) in all the settings. Regarding CoSIGN, note that since CoSIGN was specifically trained for the task of SRx4 with bicubic kernel and $\sigma_y = 0.025$, we cannot expect to outperform it in this setting (though, CM4IR yields similar PSNR). Yet, the limitations of such a task-specific approach are clearly observed in the other settings: SR with a different noise level, deblurring and inpainting, where the observation models are not identical to those assumed in [39]. The lack of flexibility of task-specific approaches is

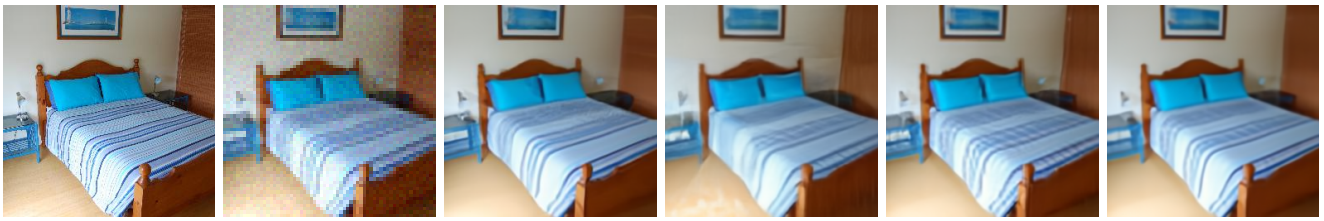

Figure 6. Super-resolution with noise level 0.025. From left to right: original, observation, DDRM(20 NFEs), DDRM(4 NFEs, auto-calculated), DDRM(4 NFEs, optimized) and DDRM(4 NFEs with our $\hat{\mathbf{z}}^-$ instead of $\hat{\mathbf{z}}$).

Table 4. Reducing NFEs for DM-based methods. PSNR [dB] (↑) and LPIPS (↓) results on LSUN Bedroom validation set.

	Method \ NFEs, $\{\tau_n\}$	20 NFEs	4 NFEs, auto-calculated	4 NFEs, optimized	4 NFEs with our $\hat{\mathbf{z}}^-$ instead of $\hat{\mathbf{z}}$
SRx4, $\sigma_y = 0.025$	DDRM	25.67 / **0.316**	24.16 / 0.395	25.40 / 0.325	**25.89** / 0.327
SRx4, $\sigma_y = 0.025$	DiffPIR	25.09 / 0.374	24.52 / 0.450	24.68 / 0.425	**25.51 / 0.371**
SRx4, $\sigma_y = 0.05$	DDRM	25.08 / **0.354**	24.12 / 0.396	24.85 / 0.361	**25.22** / 0.364
SRx4, $\sigma_y = 0.05$	DiffPIR	23.83 / 0.457	23.32 / 0.519	23.42 / 0.506	**24.89 / 0.404**

a serious limitation, as in practice, some properties of the degradation model may be unknown apriori or vary at test-time.

Importantly, we see the advantages of our CM4IR over the DM-based zero-shot methods, despite the fact that it requires significantly fewer NFEs. Specifically, CM4IR outperforms DiffPIR in all tasks, and outperforms DDRM in SR and inpainting while being competitive in deblurring.

Finally, let us state the wall-clock run-time of the zero-shot methods. Using the same hardware – Nvidia RTX 6000 Ada Generation 48GB – the average run-time for SRx4 of a single LSUN bedroom image is: 0.159 sec with the proposed CM4IR (4 NFEs), 1.539 sec with CM (40 NFEs), 0.446 sec with DDRM (20 NFEs), 0.659 sec with DiffPIR (20 NFEs). Note that: 1) we did not optimize the implementation of our method; (2) DDRM is restricted to tasks where the SVD of $\mathbf{A}$ can be computed and stored efficiently while CM4IR does not require SVD and is applicable in more cases. In the supplementary material, we also report performance advantages (better PSNR and LPIPS) of our CM4IR over DPS (1000 NFEs) [6], whose per-image run-time on our hardware is 75.461 sec.

4.3. Improving few-step guidance of DMs

A key component in our CM4IR is switching the sign of the estimated noise compared to the DDIM method. This hints that applying this modification in guided DM methods that are based on DDIM scheme, such as DDRM [15] and DiffPIR [40], can mitigate their performance drop when drastically reducing the number of the NFEs.

In this section, we explore the performance of DDRM and DiffPIR with 4 NFEs. We examine three options: 1) we let their official code set the time points (linear spacing in DDRM and quadratic spacing in DiffPIR); 2) we optimize the time points using the procedure described above (via γ and $\overline{\alpha}_N$); and 3) we flip the sign of their injected estimated noise, i.e., use our $\hat{\mathbf{z}}^-$ instead of the DDIM's original $\hat{\mathbf{z}}$ in their implementation and optimize the time points (via γ and $\overline{\alpha}_N$). We state the optimized values in the supp. mat.

The quantitative results for the SR tasks are presented in Table 4 and qualitative results are presented in Figure 6. Noticeable performance drop is observed when reducing the NFE count from 20 to 4 without modification of the noise injection (optimizing the four time points only slightly improves the results). However, utilizing our modified noise injection the performance is boosted, and quite remarkably, the PSNR becomes better than in the original implementation that uses 20 NFEs. Still, the results are inferior to those of our CM4IR, which further highlights the advantages of using CMs, rather than DMs, for few-step zero-shot restoration.

5. Conclusion

In this paper, we proposed a few-step zero-shot restoration scheme, CM4IR, that utilizes the capabilities of Consistency Models (CMs). We devised our method by combining several ingredients: initialization, back-projection guidance, and above all a novel noise injection mechanism. Considering image super-resolution, deblurring and inpainting tasks, we showed that applying our approach with as little as 4 NFEs has already advantages over alternative methods, including zero-shot methods that use many more NFEs of Diffusion Models (DMs).

Interestingly, we also showed that the usefulness of our noise injection technique is not limited to CMs. Specifically, we applied it to improve the performance of existing guided DM methods when drastically reducing NFE number. Yet, these variants did not outperform our CM4IR.

Acknowledgment

The work is supported by ISF (No. 1940/23) and MOST (No. 0007091) grants.

References

[1] Shady Abu-Hussein and Raja Giryes. Udpm: Upsampling diffusion probabilistic models. *arXiv preprint arXiv:2305.16269*, 2023. 2, 3

[2] Shady Abu Hussein, Tom Tirer, and Raja Giryes. Correction filter for single image super-resolution: Robustifying off-the-shelf deep super-resolvers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1428–1437, 2020. 2

[3] Shady Abu-Hussein, Tom Tirer, and Raja Giryes. Adir: Adaptive diffusion for image reconstruction. *arXiv preprint arXiv:2212.03221*, 2022. 2, 4

[4] Brian DO Anderson. Reverse-time diffusion equation models. *Stochastic Processes and their Applications*, 12(3):313–326, 1982. 2

[5] Ashish Bora, Ajil Jalal, Eric Price, and Alexandros G Dimakis. Compressed sensing using generative models. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 537–546. JMLR. org, 2017. 2

[6] Hyungjin Chung, Jeongsol Kim, Michael T Mccann, Marc L Klasky, and Jong Chul Ye. Diffusion posterior sampling for general noisy inverse problems. In *International Conference on Learning Representations*, 2023. 1, 2, 4, 8, 13

[7] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. *IEEE transactions on pattern analysis and machine intelligence*, 38(2):295–307, 2015. 2

[8] Bradley Efron. Tweedie's formula and selection bias. *Journal of the American Statistical Association*, 106(496):1602–1614, 2011. 3

[9] Tomer Garber and Tom Tirer. Image restoration by denoising diffusion models with iteratively preconditioned guidance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 25245–25254, 2024. 2, 4, 6, 7, 13

[10] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014. 2

[11] Magnus Rudolph Hestenes, Eduard Stiefel, et al. *Methods of conjugate gradients for solving linear systems*. NBS Washington, DC, 1952. 4

[12] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 2, 3, 5, 6

[13] Shady Abu Hussein, Tom Tirer, and Raja Giryes. Image-adaptive gan based reconstruction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3121–3129, 2020. 2

[14] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *Advances in Neural Information Processing Systems*, 35:26565–26577, 2022. 2, 3

[15] Bahjat Kawar, Michael Elad, Stefano Ermon, and Jiaming Song. Denoising diffusion restoration models. *Advances in Neural Information Processing Systems*, 35:23593–23606, 2022. 1, 2, 4, 5, 6, 7, 8

[16] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1097–1105, 2012. 2

[17] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 136–144, 2017. 2

[18] Boris T Polyak. Some methods of speeding up the convergence of iteration methods. *Ussr computational mathematics and mathematical physics*, 4(5):1–17, 1964. 2, 5

[19] Boris T Polyak. Introduction to optimization. optimization software. *Inc., Publications Division, New York*, 1(32):1, 1987. 5

[20] Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. *arXiv preprint arXiv:2202.00512*, 2022. 2, 3

[21] Assaf Shocher, Nadav Cohen, and Michal Irani. "zero-shot" super-resolution using deep internal learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3118–3126, 2018. 2

[22] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021. 2, 3, 5, 6, 11

[23] Jiaming Song, Arash Vahdat, Morteza Mardani, and Jan Kautz. Pseudoinverse-guided diffusion models for inverse problems. In *International Conference on Learning Representations*, 2023. 2, 4

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

[25] Yang Song, Liyue Shen, Lei Xing, and Stefano Ermon. Solving inverse problems in medical imaging with score-based generative models. *arXiv preprint arXiv:2111.08005*, 2021. 2, 4

[26] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021. 2, 3

[27] Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. In *Proceedings of the 40th International Conference on Machine Learning*, pages 32211–32252. PMLR, 2023. 2, 3, 4, 6, 7

[28] Tom Tirer and Raja Giryes. Image restoration by iterative denoising and backward projections. *IEEE Transactions on Image Processing*, 28(3):1220–1234, 2018. 2, 4, 13

[29] Tom Tirer and Raja Giryes. Super-resolution via image-adapted denoising CNNs: Incorporating external and internal learning. *IEEE Signal Processing Letters*, 26(7):1080–1084, 2019. 2

[30] Tom Tirer and Raja Giryes. Back-projection based fidelity term for ill-posed linear inverse problems. *IEEE Transactions on Image Processing*, 29(1):6164–6179, 2020. 4

[31] Tom Tirer and Raja Giryes. On the convergence rate of projected gradient descent for a back-projection based objective. *SIAM Journal on Imaging Sciences*, 14(4):1504–1531, 2021. 4

[32] Tom Tirer, Raja Giryes, Se Young Chun, and Yonina C Eldar. Deep internal learning: Deep learning from a single input. *IEEE Signal Processing Magazine*, 41(4):40–57, 2024. 2

[33] Singanallur V Venkatakrishnan, Charles A Bouman, and Brendt Wohlberg. Plug-and-play priors for model based reconstruction. In *2013 IEEE Global Conference on Signal and Information Processing*, pages 945–948. IEEE, 2013. 2, 4

[34] Yinhuai Wang, Jiwen Yu, and Jian Zhang. Zero-shot image restoration using denoising diffusion null-space model. In *International Conference on Learning Representations*, 2023. 2, 4, 13

[35] Kai Zhang, Wangmeng Zuo, Yunjin Chen, Deyu Meng, and Lei Zhang. Beyond a Gaussian denoiser: Residual learning of deep cnn for image denoising. *IEEE Transactions on Image Processing*, 26(7):3142–3155, 2017. 2

[36] Kai Zhang, Wangmeng Zuo, Shuhang Gu, and Lei Zhang. Learning deep cnn denoiser prior for image restoration. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3929–3938, 2017. 2, 4

[37] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023. 4

[38] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 6

[39] Jiankun Zhao, Bowen Song, and Liyue Shen. Cosign: Few-step guidance of consistency model to solve general inverse problems. In *The European Conference on Computer Vision (ECCV)*, 2024. 2, 4, 6, 7

[40] Yuanzhi Zhu, Kai Zhang, Jingyun Liang, Jiezhang Cao, Bihan Wen, Radu Timofte, and Luc Van Gool. Denoising diffusion models for plug-and-play image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1219–1229, 2023. 1, 2, 4, 5, 7, 8