A Modular Conditional Diffusion Framework for Image Reconstruction

Magauiya Zhussip[*‡]
MTS AI
m.zhussip@mts.ai

Iaroslav Koshelev[*]
AI Foundation and Algorithm Lab
ys.koshelev@gmail.com

Stamatis Lefkimmiatis[‡]
MTS AI
s.lefkimmiatis@mts.ai

Abstract

Diffusion Probabilistic Models (DPMs) have been recently utilized to deal with various blind image restoration (IR) tasks, where they have demonstrated outstanding performance in terms of perceptual quality. However, the task-specific nature of existing solutions and the excessive computational costs related to their training, make such models impractical and challenging to use for different IR tasks than those that were initially trained for. This hinders their wider adoption, especially by those who lack access to powerful computational resources and vast amount of training data. In this work we aim to address the above issues and enable the successful adoption of DPMs in practical IR-related applications. Towards this goal, we propose a modular diffusion probabilistic IR framework (DP-IR), which allows us to combine the performance benefits of existing pre-trained state-of-the-art IR networks and generative DPMs, while it requires only the additional training of a relatively small module (0.7M params) related to the particular IR task of interest. Moreover, the architecture of the proposed framework allows for a sampling strategy that leads to at least four times reduction of neural function evaluations without suffering any performance loss, while it can also be combined with existing acceleration techniques such as DDIM. We evaluate our model on four benchmarks for the tasks of burst JDD-SR, dynamic scene deblurring, and super-resolution. Our method outperforms existing approaches in terms of perceptual quality while it retains a competitive performance with respect to fidelity metrics.

1 Introduction

With the advent of deep learning we have witnessed outstanding results in a wide range of computer vision tasks [89], including many challenging blind image restoration (IR) problems [84] such as burst imaging [40], super-resolution (SR) [9], deconvolution [58], *etc*. The standard approach for supervised learning in a blind IR setting involves training a feed-forward network that should estimate the latent image based on the available low-quality measurements. Such models are usually trained to maximize *fidelity* metrics like PSNR or SSIM, but the visual quality of the resulting images is sub-optimal [6]. The inclusion of perceptual losses [30] to the objective can improve the visual results, but fails to convincingly address the problem.

A promising direction towards IR results of high visual quality is to consider such problems within a generative framework. Several generative models have been recently proposed including Variational Autoencoders (VAEs) [33], Generative Adversarial Neural Networks (GANs) [22], Normalizing Flows (NFs) [16] and Diffusion Probabilistic Models (DPMs) [65]. Due to their impressive results in image generation, they have been further utilized to perform *conditional sampling* of high-quality images, with their low-quality or distorted counterparts playing the role of the conditional

[*]Equal contribution
[‡]Work performed while at AI Foundation and Algorithm Lab

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

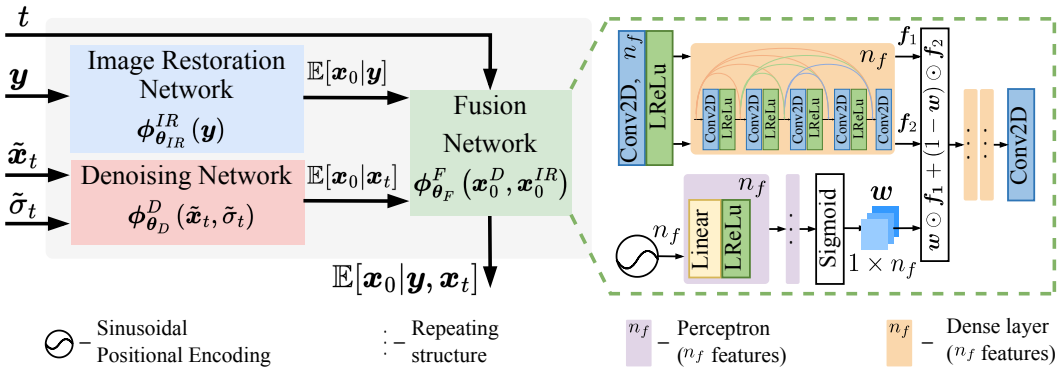

Figure 1: The proposed architecture consists of three modules: a Denoising Network $\phi^{D}_{\theta_D}(\tilde{\boldsymbol{x}}_t, \tilde{\sigma}_t)$, an IR Network $\phi^{IR}_{\theta_{IR}}(\boldsymbol{y})$ and a Fusion Network $\phi^{F}_{\theta_F}(\boldsymbol{x}^{IR}_0, \boldsymbol{x}^{D}_0, t)$. A small version of MIRNet [81] is used as the Denoising Network, while a pre-trained SwinIR [42] or BSRT [50] or FFTFormer [34] is used as the IR Network, depending on the IR task. See section 3.3 for a detailed description.

input [21, 38, 46, 39]. To date, DPMs appear to be the most promising framework and lead to the best results among all existing generative approaches.

Nevertheless, there are certain limitations that existing DPMs face, which hinder their wider adoption in IR tasks. In particular, inference of such models involves a sampling process that requires a large number (in the order of hundreds) of neural function evaluations (NFEs), which can be computationally very expensive, especially when considering images of high resolution. Another important limitation is that an efficient conditioning on the image measurements has yet to be proposed for DPMs in order to make them applicable to a wider range of blind IR problems. Indeed, all of the existing methods aim to learn the parameters of a single input-conditioned network for a specific blind IR task. As a result, the trained model overfits on the distribution of the condition space, and the whole model has to be retrained if we need to employ it to a different reconstruction task than the one that was initially trained for. Considering the huge amount of data and computational resources required for training a single DPM (see appendix A), such re-training becomes infeasible if at least one of the previous requirements is not satisfied.

In this work we aim to address the above issues by proposing a novel conditional diffusion network coupled with an accelerated sampling process. Specifically, our network adopts an improved conditioning strategy and is built on the foundation of existing off-the-shelf IR networks paired with a denoising module, which is applicable to a variety of reconstruction problems without requiring any re-training. Additionally, we introduce an accelerated sampling procedure that is enabled by our proposed network architecture and allows the merging of a large number of sampling steps in a single one, computed with a single NFE. Our proposed acceleration can work in tandem with accelerated sampling schemes such as DDIM [66]. To assess the performance of our network, we validate it on three challenging blind IR tasks, namely, burst joint demosaicking, denoising and super-resolution (JDD-SR), dynamic scene deblurring, and $4\times$ single image super-resolution (SISR). In all of the tested scenarios, our approach demonstrates the best perception-distortion trade-off among the state-of-the-art (SOTA) methods, while compared to other DPM-based solutions it requires a smaller number of sampling steps.

2 Related Works

Burst Image Restoration. One of the pioneering works in multi-frame IR was introduced in [70], where a frequency-domain-based solution was proposed. Then, several MAP models with various regularization terms have been designed to cope with visual artifacts caused by operating in the frequency domain [3, 18, 63]. Using the same MAP framework, a JDD-SR method robust to noise and outliers was developed in [19]. Meanwhile, the block matching alignment algorithm of [25] was extended by [79] to obtain a robust motion model with the aid of an adaptive kernel interpolation method merging sparse pixel samples.

Advancements in deep learning have led to high-performing methods such as those in [17, 37, 4, 5, 49, 41]. The DBSR approach [4] aligns multiple input frames in the feature space utilizing an optical flow estimator (*e.g.* PWCNet [68]) and employs an attention-based fusion mechanism to aggregate

features. In [37] a differentiable image registration module has been introduced, which exploits the aliasing effects appearing in bursts of low-resolution (LR) images. In [41] KBNet estimates blur kernels for a burst sequence to incorporate them with LR features so as to generate a better super-resolved image, while in [17] BIPNet attempts to fuse complementary information from the burst sequence with the help of generated pseudo-burst features. Another line of work effectively employs deformable convolutions for the inter-frame alignment task [49, 74, 17] and achieves SOTA results in various tasks, including burst SR.

Single Image Restoration. Among the recent IR methods, the most successful ones are those that adopt an end-to-end supervised formulation, where a deep neural network is trained to directly map a low-quality and degraded image to a point estimate of the latent high-quality image [88, 45, 69, 83]. Consequently, in the pursuit of further improving the reconstructions and achieving a better pixel-level result, more advanced network architectures have been proposed [82, 42, 10, 34], at the cost of being more computationally heavy. While this formulation leads to SOTA fidelity (*e.g.* , PSNR, SSIM), the produced output is an average/median of all plausible predictions, which typically lacks high-frequency information (*e.g.* texture).

Generative adversarial networks (GAN) [22] have been adopted by several IR methods such as SISR [38, 73, 75] and dynamic scene deblurring [35, 36, 85] to produce more natural and perceptually pleasing results. Although this adversarial non-reference formulation aims to push the predictions towards the manifold of natural images, it is also prone to introducing unrealistic texture and hallucinations in the output [14]. Moreover, the adversarial training process requires extra supervision as it can easily fall into a mode collapse or may diverge [2, 62].

Likelihood-based deep generative models such as NFs [47, 46], auto-regressive models [23], and VAEs [57] have also been applied to IR tasks, where one can obtain a diverse set of predictions from a learned posterior [57]. Conditioned on LR inputs, flow-based methods attempt to map high-resolution (HR) images to the latent flow-space. Although such techniques circumvent the training instability met in GANs, strong architectural constraints (*e.g.* network invertibility) still remain an issue.

Recently another class of methods based on a stochastic diffusion process has been introduced and demonstrated outstanding performance on various tasks that range from unconditional image generation [26, 55, 60] to image-to-image translation/restoration [39, 61, 78, 20, 60, 15, 76, 59]. DvSR proposed in [78] employs a "predict-and-refine" conditional diffusion method specifically tailored for the image deblurring task, while SRDiff [39] utilizes features of a pretrained SR model for conditional super-resolved image generation. Further, recent works in [76, 80] have considered several IR tasks (*e.g.* inpainting, super-resolution, colorization, etc.). In conclusion, their ability to capture complex statistics of the visual world, makes DPMs a very attractive solution that is worth being further investigated.

3 Proposed Conditional Diffusion Model

3.1 Background

Denoising Diffusion Probabilistic Models (DDPMs) [26, 65] are special cases of Hierarchical Markovian Variational Autoencoders where the dimension of the latent variables matches the dimension of the data. Starting with a sample $\boldsymbol{x}_0 \in \mathbb{R}^N$, the encoding sequence $\{\boldsymbol{x}_t\}_{t=0}^T$ traverses the latent space with a *diffusion process* defined by a Gaussian transition probability:

$$q\left(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}\right) \equiv \mathcal{N}\left(\boldsymbol{x}_t; \sqrt{1-\beta_t}\boldsymbol{x}_{t-1}, \beta_t \boldsymbol{I}_N\right). \tag{1}$$

The sequence $0 < \beta_1, \beta_2, \ldots, \beta_T < 1$ that appears in eq. (1) defines the noise scheduling for the forward process in such a way so that the latent variable at the final timestep T approximates the standard Gaussian: $\boldsymbol{x}_T \sim \mathcal{N}\left(\boldsymbol{x}_T; \mathbf{0}, \boldsymbol{I}_N\right)$. Based on this diffusion process, it is possible to express the transition probability directly from $\boldsymbol{x}_0$ to $\boldsymbol{x}_t$ in closed form as:

$$q\left(\boldsymbol{x}_t|\boldsymbol{x}_0\right) = \mathcal{N}\left(\boldsymbol{x}_t; \sqrt{\bar{\alpha}_t}\boldsymbol{x}_0, \left(1-\bar{\alpha}_t\right)\boldsymbol{I}_N\right), \tag{2}$$

where $\alpha_t \equiv 1 - \beta_t$ and $\bar{\alpha}_t \equiv \prod_{s=1}^{t} \alpha_s$.

The *reverse process* is enabled by the posterior distribution which is represented in the form:

$$p\left(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0\right) = \mathcal{N}\left(\boldsymbol{x}_{t-1}; \boldsymbol{\mu}_t\left(\boldsymbol{x}_t, \boldsymbol{x}_0\right), \sigma_t^2 \boldsymbol{I}_N\right), \tag{3}$$

where $\boldsymbol{\mu}_t(\boldsymbol{x}_t, \boldsymbol{x}_0) \equiv \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}\boldsymbol{x}_0 + \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\boldsymbol{x}_t$ and $\sigma_t^2 \equiv \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t$. DDPMs aim to approximate its mean by the quantity $\boldsymbol{\mu}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t)$, which is learned from training data, and then utilize eq. (3) to perform sampling. There are different possible parameterizations of $\boldsymbol{\mu}_t(\boldsymbol{x}_t, \boldsymbol{x}_0)$, which accordingly lead to different interpretations for the transition mean [48]. In this work, we pursue the one based on the score function $\nabla \log p(\boldsymbol{x}_t)$, which reads as:

$$\boldsymbol{\mu}_t(\boldsymbol{x}_t, \boldsymbol{x}_0) = \frac{\boldsymbol{x}_t + (1-\alpha_t)\nabla \log p(\boldsymbol{x}_t)}{\sqrt{\alpha_t}}. \tag{4}$$

In this case, the reverse process defined in eq. (3) can be considered as sampling via Annealed Langevin Dynamics, in which the score function is approximated by the quantity $\boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t)$ learned via denoising score matching [28, 72].

3.2 Conditional Score Matching

The diffusion models described above do not take into account the dependence of the sampled data on their degraded observations $\boldsymbol{y} \in \mathbb{R}^M$, when we are dealing with IR problems. Fortunately, the score-based models can be extended to accommodate conditional sampling by replacing the score function in eq. (4) with a conditional score function $\nabla_{\boldsymbol{x}_t} \log p(\boldsymbol{x}_t|\boldsymbol{y})$. For non-blind IR problems, a popular approach is to decompose the conditional score function into a score function $\nabla \log p(\boldsymbol{x}_t)$ and a log-likelihood gradient term $\nabla_{\boldsymbol{x}_t} \log p(\boldsymbol{y}|\boldsymbol{x}_t)$ [54, 67]. This last term is directly dependent on the image formation model, which unfortunately is unknown for blind IR tasks. Therefore, most of the existing works [39, 59, 60] aim instead to learn the primal conditional score function $\nabla_{\boldsymbol{x}_t} \log p(\boldsymbol{x}_t|\boldsymbol{y})$ via *ad-hoc* conditional denoising score matching. In this work, we also utilize the primal conditional score function, but we rely on its explicit form as given in the following lemma, whose proof is provided in the appendix B.

Lemma 3.1. *Let $\boldsymbol{y} \in \mathbb{R}^M$, $\boldsymbol{x}_0 \in \mathbb{R}^N \sim p(\boldsymbol{x}_0|\boldsymbol{y})$, and $\boldsymbol{x}_t \in \mathbb{R}^N$, $\bar{\alpha}_t \in \mathbb{R}$ are defined as in eq.* (2)*. Then, the conditional score function is computed as:*

$$\nabla_{\boldsymbol{x}_t} \log p(\boldsymbol{x}_t|\boldsymbol{y}) = \frac{\sqrt{\bar{\alpha}_t}\,\mathbb{E}[\boldsymbol{x}_0|\boldsymbol{y}, \boldsymbol{x}_t] - \boldsymbol{x}_t}{1-\bar{\alpha}_t}. \tag{5}$$

The above result implies that the conditional score function can be approximated by utilizing a trained joint reconstruction and denoising model. Specifically, if the augmented variable $\boldsymbol{z}_t = \begin{bmatrix}\boldsymbol{y}^\mathsf{T} & \boldsymbol{x}_t^\mathsf{T}\end{bmatrix}^\mathsf{T}$ represents the union of the degraded data $\boldsymbol{y}$ and the noisy data $\boldsymbol{x}_t$, then the conditional expected value $\mathbb{E}[\boldsymbol{x}_0|\boldsymbol{z}_t]$ corresponds to the reconstructed underlying image $\boldsymbol{x}_0$ from the measurements $\boldsymbol{z}_t$. A joint reconstruction model $\boldsymbol{\phi}_\theta(\boldsymbol{y}, \boldsymbol{x}_t, t)$ can be trained by the minimization of the empirical expected pixel mean-squared error (MSE) across the samples from the training dataset $\mathcal{D}(\boldsymbol{x}_0, \boldsymbol{y}, \boldsymbol{x}_t, t)$:

$$\min_{\boldsymbol{\theta}} \mathbb{E}_{\boldsymbol{x}_0, \boldsymbol{x}_t, \boldsymbol{y} \sim \mathcal{D}} \|\boldsymbol{\phi}_{\boldsymbol{\theta}}(\boldsymbol{y}, \boldsymbol{x}_t, t) - \boldsymbol{x}_0\|_2^2 = \min_{\boldsymbol{\theta}} \sum_i \|\boldsymbol{\phi}_{\boldsymbol{\theta}}(\boldsymbol{y}^i, \boldsymbol{x}_t^i, t) - \boldsymbol{x}_0^i\|_2^2. \tag{6}$$

The optimal solution is the conditional expectation $\boldsymbol{\phi}_\theta^{\text{MSE}}(\boldsymbol{y}, \boldsymbol{x}_t, t) = \mathbb{E}[\boldsymbol{x}_0|\boldsymbol{y}, \boldsymbol{x}_t]$, and, thus, such a trained model can be substituted in eq. (5). This amounts to approximating the conditional score function $\nabla_{\boldsymbol{x}_t} \log p(\boldsymbol{x}_t|\boldsymbol{y})$ with $\boldsymbol{s}_{\boldsymbol{\theta}}^c(\boldsymbol{y}, \boldsymbol{x}_t, t) \equiv \frac{\sqrt{\bar{\alpha}_t}\boldsymbol{\phi}_\theta(\boldsymbol{y}, \boldsymbol{x}_t, t) - \boldsymbol{x}_t}{1-\bar{\alpha}_t}$.

3.3 Proposed Network Architecture

Our objective is to parameterize the function $\boldsymbol{\phi}_\theta(\boldsymbol{y}, \boldsymbol{x}_t, t)$ in a form of a neural network (CNN) and design a specific architecture of this network. The absence of explicit knowledge about the formation model $\boldsymbol{x}_0 \to \boldsymbol{y}$ requires the network to learn it implicitly from training data. Such an approach generally results in over-fitting, meaning that the trained model can only be employed for the task it was originally trained for [39, 78]. To overcome this problem, we initially build on the hypothesis that the conditional expectation $\mathbb{E}[\boldsymbol{x}_0|\boldsymbol{y}, \boldsymbol{x}_t]$ related to the conditional score function in eq. (5), can be approximated by a function of two easier to compute conditional expectations, that is

$$\mathbb{E}[\boldsymbol{x}_0|\boldsymbol{y}, \boldsymbol{x}_t] \approx f(\mathbb{E}[\boldsymbol{x}_0|\boldsymbol{y}], \mathbb{E}[\boldsymbol{x}_0|\boldsymbol{x}_t]). \tag{7}$$

Based on such an approximation, it is now possible to learn a single unconditional generative denoising model that can be applied in different reconstruction problems. Further, we note that despite the absence of a good approximation of the likelihood term, $\mathbb{E}[\boldsymbol{x}_0|\boldsymbol{y}]$, various task-specific networks trained with a fidelity objective are readily available in the literature. Indeed, using a

similar reasoning as the one provided for eq. (6), such reconstruction networks can output a good approximation of the quantity $\mathbb{E}[\boldsymbol{x}_0|\boldsymbol{y}]$. This finally motivates us to express the joint reconstruction and denoising network $\boldsymbol{\phi}_o(\boldsymbol{y}, \boldsymbol{x}_t, t)$ into three components (see Figure 1). Specifically, our network can be described as $\boldsymbol{\phi}^F_{\boldsymbol{\theta}_F}\left(\boldsymbol{\phi}^{IR}_{\boldsymbol{\theta}_{IR}}(\boldsymbol{y}), \boldsymbol{\phi}^D_{\boldsymbol{\theta}_D}(\tilde{\boldsymbol{x}}_t, \tilde{\sigma}_t), t\right)$, where $\tilde{\boldsymbol{x}}_t \equiv \frac{\boldsymbol{x}_t}{\sqrt{\bar{\alpha}_t}} \sim \mathcal{N}\left(\boldsymbol{x}_0, \tilde{\sigma}_t^2 \boldsymbol{I}_N\right)$ is the noisy version of $\boldsymbol{x}_0$ with noise variance $\tilde{\sigma}_t^2 \equiv \frac{1-\bar{\alpha}_t}{\bar{\alpha}_t}$ according to eq. (2), and the sub-modules $\boldsymbol{\phi}^D_{\boldsymbol{\theta}_D}, \boldsymbol{\phi}^{IR}_{\boldsymbol{\theta}_{IR}}, \boldsymbol{\phi}^F_{\boldsymbol{\theta}_F}$ are defined next.

IR network $\boldsymbol{\phi}^{IR}_{\boldsymbol{\theta}_{IR}}(\boldsymbol{y})$, which is learned in a supervised manner to predict $\mathbb{E}[\boldsymbol{x}_0|\boldsymbol{y}]$. Specifically, we employ the BSRT-Small [50] for burst JDD-SR, FFTformer [34] for dynamic scene deblurring, and SwinIR [42] for SISR. We do not train these networks but use their publicly available trained weights.

Denoising network $\boldsymbol{\phi}^D_{\boldsymbol{\theta}_D}(\tilde{\boldsymbol{x}}_t, \tilde{\sigma}_t)$, which is learned in a supervised manner to predict $\mathbb{E}[\boldsymbol{x}_0|\boldsymbol{x}_t]$ by denoising $\tilde{\boldsymbol{x}}_t$. Specifically, we employ a smaller version of MIRNet [81], which we call MIRNet-S. It is obtained by reducing the amount of RRG and MRB blocks from the original architecture to three and one, respectively. We refer to the original paper [81] for the detailed description of the blocks structure, as we use them without any modifications. Once trained, this network is reused for all considered reconstruction problems. We note that we our motivation for utilizing a smaller version of MIRNet as a Denoising module, is to approximately match the number of parameters and the computational complexity of the networks used in our framework with those of the alternative methods under study. This way we can ensure a fair evaluation and comparison among competing methods. Such strategy has allowed us to achieve direct performance comparisons under similar conditions.

Fusion network $\boldsymbol{\phi}^F_{\boldsymbol{\theta}_F}\left(\boldsymbol{x}^{IR}_0, \boldsymbol{x}^D_0, t\right)$, which predicts the conditional expectation $\mathbb{E}[\boldsymbol{x}_0|\boldsymbol{y}, \boldsymbol{x}_t]$. This module refines and combines the predictions of the previous two networks and is the only one that needs to be trained for each specific IR task. The fusion network accepts as inputs the image estimates $\boldsymbol{x}^{IR}_0, \boldsymbol{x}^D_0$ and a timestep t. Its architecture consists of two branches. The first one involves a convolution layer with n_f output channels followed by a single dense block [27] without batch normalization. Its purpose is to independently encode both input images into the corresponding features $\boldsymbol{f}_1, \boldsymbol{f}_2$ with n_f channels each. The second branch encodes the timestep t into a vector of weights $\boldsymbol{w} \in (0,1)^{n_f}$ using the sinusoidal positional encoding [71], followed by a two layer perceptron and a sigmoid function as the final activation. The features $\boldsymbol{f}_1, \boldsymbol{f}_2$ and the weights $\boldsymbol{w}$ are then passed to the Convex Combination Channel Attention (3CA) layer, which performs the per-channel aggregation of input features as a convex combination of the form: $\boldsymbol{w} \odot \boldsymbol{f}_1 + (\boldsymbol{1} - \boldsymbol{w}) \odot \boldsymbol{f}_2$. The output of this layer is decoded by two consequent dense blocks with n_f channels each, followed by a convolution layer which produces the final output $\mathbb{E}[\boldsymbol{x}_0|\boldsymbol{y}, \boldsymbol{x}_t]$. This proposed architecture results in a significantly smaller network size than those of the Denoising and IR modules. Thus we can train the fusion network fast and by using only a small amount of problem-specific training data. While we explored several basic fusion architectures, we did not delve into extensive research to ascertain the optimal design. Our proposed fusion module serves as a proof of concept, validating our framework and demonstrating its potential for performance enhancement. A comprehensive investigation into optimal fusion architectures remains a promising area for future research.

Such a modular overall architecture allows us to capitalize on the existing SOTA non-blind denoising and blind IR networks, while it also allows us to easily replace any of these networks when better ones become available in the future. As we describe next, another important advantage of our proposed pipeline, is that it allows us to achieve a significant acceleration for the sampling process without incurring any loss of reconstruction quality.

3.4 Proposed Accelerated Sampling

According to eq. (3), our conditional denoiser should be evaluated for all timesteps $t = \overline{T, ..., 0}$, which leads to a total of T NFEs. We note that by construction, for the forward process it holds that $\boldsymbol{x}_T \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_N)$. This means that in the beginning of sampling, the latent variable $\boldsymbol{x}_T$ does not contain any information about $\boldsymbol{x}_0$. It is also reasonable to expect that a similar lack of information about $\boldsymbol{x}_0$ exists for a number of steps prior to T. Specifically, for those steps we expect that the quantity $\mathbb{E}[\boldsymbol{x}_0|\boldsymbol{y}, \boldsymbol{x}_t]$ is heavily influenced by $\mathbb{E}[\boldsymbol{x}_0|\boldsymbol{y}]$, while the contribution of $\mathbb{E}[\boldsymbol{x}_0|\boldsymbol{x}_t]$ is not significant enough. A theoretical justification for this argument is provided in appendix C. Based on the above reasoning, we select a timestep τ such that for the first $T - \tau$ reverse steps we use the following approximation: $\mathbb{E}[\boldsymbol{x}_0|\boldsymbol{y}, \boldsymbol{x}_t] \approx \mathbb{E}[\boldsymbol{x}_0|\boldsymbol{y}] = \boldsymbol{\phi}^{IR}_{\boldsymbol{\theta}_{IR}}(\boldsymbol{y})$. This is achieved by disabling the lower branch of our proposed conditional score matching network, namely the Denoising $\boldsymbol{\phi}^D_{\boldsymbol{\theta}_D}(\tilde{\boldsymbol{x}}_t, \tilde{\sigma}_t)$ and Fusion $\boldsymbol{\phi}^F_{\boldsymbol{\theta}_F}\left(\boldsymbol{x}^{IR}_0, \boldsymbol{x}^D_0\right)$ modules (Figure 1). Our strategy can be further supported by the recent

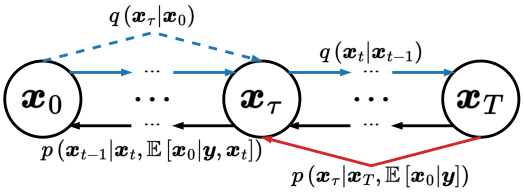

Figure 2: Forward and reverse diffusion process. Blue solid arrows: transitions at the forward pass with sampling distribution from eq. (1). Dashed arrow: cumulative transition probability from eq. (2). Black solid arrows: transitions at the backward pass with the sampling distribution from eq. (3). Red solid arrow: closed-form cumulative transition probability from eq. (8) representing our accelerated sampling.

Table 1: Performance evaluation on the task of Burst JDD-SR. We highlight the overall best for each metric.

Methods	PSNR↑	SSIM↑	LPIPS↓	TOPIQ$_\Delta$↓	NFE ↓	Params↓
Target	∞	1	0	0	N/A	N/A
DBSR	31.98	0.891	0.198	0.10	N/A	13.0M
DeepRep	34.66	0.927	0.136	0.07	N/A	12.1M
EBSR	36.05	0.940	0.111	0.15	N/A	26.0M
BIPNet	34.86	0.934	0.112	0.03	N/A	6.7M
BSRT-Small	35.91	0.940	0.109	0.12	N/A	4.9M
BSRT-Large	36.98	0.947	0.095	0.16	N/A	20.7M
Ours	35.53	0.933	0.084	0.02	6	21.6M

study in [12], where it has been demonstrated that the image sampling via DPMs could be divided into stages depending on the reverse process timesteps. In this spirit we activate the Denoising and Fusion modules at a timestep τ that is selected experimentally for the particular IR task of interest. Our results clearly indicate that the reconstruction result is going to be exactly the same whether we utilize the multi-step reverse diffusion process or the proposed one-step process that is described in Lemma 3.2. Indeed, since the quantity $\mathbb{E}[\boldsymbol{x}_0|\boldsymbol{y}]$ is predicted by the IR network, which does not depend on the reverse diffusion parameters, $\phi^{IR}_{\boldsymbol{\theta}_{IR}}(\boldsymbol{y})$ needs to be evaluated only once and its output can be re-used throughout the whole iterative sampling procedure. Expanding more on this idea, we show in Lemma 3.2 that it is possible to omit entirely the first $T-\tau$ reverse diffusion steps and instead perform a single step directly from T to τ with a procedure very similar to the one obtained for the diffusion process in eq. (2) from eq. (1). We provide the derivation in appendix D.

Lemma 3.2. *The transition probability defined in eq.* (3) *for a single reverse step, can be extended to* k *reverse steps starting from* $\boldsymbol{x}_t$ *as:*

$$p(\boldsymbol{x}_{t-k}|\boldsymbol{x}_t, \boldsymbol{x}_0) = \mathcal{N}\left(\boldsymbol{x}_{t-k}; \boldsymbol{\mu}_{t,k}(\boldsymbol{x}_t, \boldsymbol{x}_0), \sigma^2_{t,k}\boldsymbol{I}_N\right), \tag{8}$$

where

$$\boldsymbol{\mu}_{t,k}(\boldsymbol{x}_t, \boldsymbol{x}_0) = \sum_{i=0}^{k-1} \Gamma^{t-k+1}_{t-i-1}\gamma_{t-i}\boldsymbol{x}_0 + \Gamma^{t-k+1}_t \boldsymbol{x}_t \;\; \textit{and} \;\; \sigma^2_{t,k} = \sum_{i=0}^{k-1}\left(\Gamma^{t-k+1}_{t-i-1}\right)^2\sigma^2_{t-i}. \tag{9}$$

In the above equations we make use of the following notation:

$$\gamma_t = \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t} \;\; \textit{and} \;\; \Gamma^j_i \equiv \begin{cases} \sqrt{\prod_{n=j}^{i}\alpha_n}\frac{1-\bar{\alpha}_{j-1}}{1-\bar{\alpha}_i} & \textit{for } i \geq j \\ 1 & \textit{for } i < j \end{cases} \tag{10}$$

Since for the first $T-\tau$ steps $\boldsymbol{x}_0$ is approximated by the quantity $\mathbb{E}[\boldsymbol{x}_0|\boldsymbol{y}]$, which is independent of the timestep t, we utilize eq. (8) to directly sample $\boldsymbol{x}_\tau$ as

$$\boldsymbol{x}_\tau \sim p(\boldsymbol{x}_\tau|\boldsymbol{x}_T, \mathbb{E}[\boldsymbol{x}_0|\boldsymbol{y}]) = \mathcal{N}\left(\boldsymbol{x}_\tau; \boldsymbol{\mu}_{T,T-\tau}(\boldsymbol{x}_T, \mathbb{E}[\boldsymbol{x}_0|\boldsymbol{y}]), \sigma^2_{T,T-\tau}\boldsymbol{I}\right). \tag{11}$$

This allows us to reduce the NFEs from $T+1$ to $\tau+1$, meaning that the required evaluations of Denoising + Fusion networks is reduced from T to τ. In both cases, we additionally count a single evaluation of the IR network. We depict our acceleration strategy with a red arrow in Figure 2. Finally, in practice we use $\tau = \frac{T}{200}$ for burst JDD-SR and dynamic scene deblurring, and $\tau = \frac{T}{4}$ for SISR, effectively reducing the NFEs by two orders of magnitude and a factor of four, respectively.

The proposed acceleration procedure can be also interpreted as starting the sampling from step τ of the latent space using a non-standard Gaussian distribution as defined in eq. (11), instead of starting from step T and using a standard Gaussian sample $\boldsymbol{x}_T \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_N)$. We note, that a similar idea was explored in [13, 52], where it was proposed to start the sampling from an observation that has been passed through a predefined number of forward diffusion steps. In our case the starting point for sampling is obtained via the approximated reverse process, which as a consequence of Lemma 3.2 does not alter the final reconstruction result. In another words, if we approximate $\boldsymbol{x}_0$ with $\mathbb{E}[\boldsymbol{x}_0|\boldsymbol{y}]$, then the reconstruction result will be the same both for the multi-step reverse diffusion process and

the proposed one-step process. Moreover, it is easy to show that our strategy generalizes the one proposed in [13, 52], as it leads to the same starting point if we make the following specific choices: $\boldsymbol{x}_T \sim \mathcal{N}\left(\sqrt{\bar{\alpha}_T}\mathbb{E}\left[\boldsymbol{x}_0|\boldsymbol{y}\right], \left(1-\bar{\alpha}_T\right)\boldsymbol{I}\right)$ and $\sigma_t^2 = \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t$. In practice $\bar{\alpha}_T \approx 0$, so the first condition holds almost exactly. The second condition represents the particular choice of the noise variance used in the reverse process, with several existing parametrizations [26, 55]. Our method is compatible with all of them and leads to different distributions for the starting point. In all our experiments we use the parametrization $\sigma_t^2 = \beta_t$ from [26]. More detailed conceptual and technical differences along with experimental results are provided in appendix E. Furthermore, our acceleration strategy is complimentary to other sampling acceleration techniques [66, 31]. To demonstrate this, we utilize the DDIM [66] sampling to further reduce the NFEs by a factor of five for the SISR task. As a result, the reverse diffusion process used for this problem requires $\frac{T}{20} + 1$ NFEs in total.

4 Experimental Results

We evaluate our method on four public datasets across a range of tasks, namely burst JDD-SR, dynamic scene deblurring, and SISR. Below we describe our specific architecture and design choices related to all utilized modules.

Training. Our training procedure consists of two stages. We first employ a diverse, yet small DF2K (combination of DIV2K [1] and Flickr2K [45]) dataset to train a Denoising Module for Gaussian denoising in the sRGB domain with input noise levels ranging in $[0, 244.3]$. These noise levels corresponds to timesteps in the range of $[0, 250]$ for the diffusion process with $T = 1000$. We use the original training procedure of MIRNet [81] to learn the parameters of our MIRNet-S architecture. At the second stage we train our Fusion Modules with $n_f = 64$ for each IR task and the corresponding pre-trained off-the-shelf IR network. It is worth noting, that at this stage the parameters of Denoising and IR modules are kept frozen and only the Fusion Module is trained. Specifically, we train it for $300k$ iterations with a learning rate of $10^{-4} \times 0.99$ it/1000, batch size of 128, and crop size of 256×256. To train our Fusion networks we use datasets that are common among our main competitors, specifically the ZurichRAW2RGB [29] dataset for burst JDD-SR, GoPro [53] for dynamic scene deblurring and DIV2K [1] for $4\times$ SISR. For burst JDD-SR the Fusion network is trained in the sRGB domain. All the networks are trained using the Ascend 910 AI accelerators [44]. To make our results reproducible, we provide a full description of the training procedure in appendix H.

Inference. For each IR task we use the procedure described in section 3.4 to obtain the reconstructed images with $T = 1000$. To demonstrate the effectiveness of our approach, for each problem of interest except for burst JDD-SR we need half of the NFEs compared to the diffusion-based competitor that uses the least number of sampling steps. Specifically, for dynamic scene deblurring we use $\tau = 5$, resulting in $200\times$ acceleration achieved solely by our proposed sampling strategy. This amounts to 6 NFEs when counting the additional IR network evaluation performed to skip the first $T - \tau = 995$ steps using eq. (8). For SISR we select $\tau = 250$, which results in $4\times$ acceleration using our sampling procedure. In order to demonstrate how it can be complemented by other proposed acceleration strategies, for the final $\tau = 250$ steps we achieve $5\times$ step reduction by employing DDIM sampling [66]. The combination of both acceleration strategies results in $20\times$ step reduction and 51 NFEs overall. Applying the DDIM acceleration technique on top of our proposed one-step strategy leads to an insignificant quantitative/qualitative difference (see appendix G) compared to our original scheme. Since for the burst JDD-SR problem no diffusion-based methods have yet been proposed, we use the same setting as for the dynamic scene deblurring problem, as it requires the smallest NFEs. In all our experiments we use the linear scheduling of the diffusion process variances $\beta_t \in \left[2 \times 10^{-2}, 10^{-4}\right]$ defined in eq. (1).

Evaluation. For the burst JDD-SR evaluation we use the SyntheticBurst test set [4], consisting of 300 synthetically pre-generated raw burst sequences. Each sequence contains 14 noisy raw LR images with handshake motion, whose corresponding targets have a resolution of 320×320. Since our networks are trained on sRGB images, the outputs of all methods are converted to the sRGB space prior to comparison. For dynamic scene deblurring we evaluate on the GoPro test [53] and HIDE [64] benchmarks, which contain 1111 and 2025 images of 720p resolution, respectively. For $4\times$ SISR we use the DIV2K validation dataset [1] consisting of 100 images of 2K resolution.

For the quantitative evaluation of the reconstruction quality we rely on the widely used fidelity metrics PSNR and SSIM [77], and the reference-based perceptual metric LPIPS [86]. Moreover, we

Table 2: Performance evaluation on the GoPro and HIDE test sets for dynamic scene deblurring. † indicates that public implementation is unavailable and the scores are copied from the authors' paper. We highlight the overall best for each metric, and the best among perceptual-oriented methods.

Methods	GoPro				HIDE				NFE ↓	Params ↓
	PSNR↑	SSIM↑	LPIPS↓	TOPIQ$_\Delta$↓	PSNR↑	SSIM↑	LPIPS↓	TOPIQ$_\Delta$↓		
Target	∞	1	0	0	∞	1	0	0	N/A	N/A
HINet	32.77	0.960	0.088	0.033	30.33	0.932	0.120	0.044	N/A	88.6M
MPRNet	32.66	0.959	0.089	0.027	30.96	0.939	0.114	0.059	N/A	20.1M
MIMO-UNet+	32.44	0.957	0.091	0.034	29.99	0.930	0.124	0.028	N/A	16.1M
NAFNet	33.71	0.967	0.078	0.017	31.32	0.943	0.103	0.024	N/A	67.9M
Restormer	32.90	0.961	0.084	0.018	31.20	0.942	0.109	0.048	N/A	26.1M
FFTFormer	34.21	0.969	0.071	0.012	31.62	0.946	0.096	0.006	N/A	16.6M
Perceptual-oriented Methods										
DeblurGANv2	29.08	0.918	0.117	0.025	27.51	0.885	0.159	0.065	N/A	60.9M
DvSR†	31.66	0.948	0.059	-	29.77	0.922	0.089	-	500	26.1M
icDPM†	31.19	0.943	0.057	-	29.14	0.910	0.088	-	500	52.0M
InDi†	31.49	0.946	0.058	-			-		10	27.7M
Ours	33.72	0.963	0.053	0.011	31.32	0.937	0.087	0.002	6	33.2M

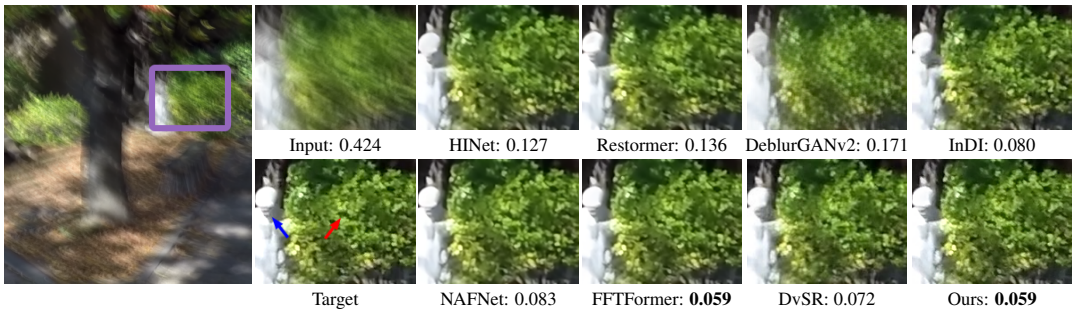

Figure 3: Visual comparisons on the GoPro test set for the task of dynamic scene deblurring (best viewed by zooming in). Every output image is accompanied by its LPIPS value.

also utilize the non-reference image quality assessment (NR-IQA) metric TOPIQ [8] and report the absolute distance between the output and the target scores, which we indicate as TOPIQ$_\Delta$.

4.1 Results

Burst JDD-SR. In Table 1 we compare our proposed pipeline with existing methods, namely DBSR [4], DeepRep [5], and the current SOTA methods, namely BIPNet [17], BSRT-Small [50], BSRT-Large [50], and EBSR [49]. Our method demonstrates SOTA performance across the perceptual metrics while maintaining competitive PSNR and SSIM scores compared to existing methods. Thus, our method reconstructs images that are closer to the target based on the human perception while maintaining a high level of fidelity. We refer to figures in the appendix M for a qualitative visual assessment. Furthermore, we notice an improvement in terms of visual quality and perceptual metrics compared to BSRT-Small, which we use as the IR module of choice in our framework. This indicates that our approach preserves the fidelity of the IR model outputs, while enhancing their perceptual quality by running few reverse diffusion steps. It is worth noting that for this case, where a burst of raw images serves as input, we use the exact same denoising network that was trained on sRGB images and which we later deploy to all considered single-input IR tasks. This highlights the generalization ability of our approach not only to different IR problems but also to different input formats. Note that AFCNet [51], LKR [37], and Burstormer [17] are not included in our comparisons due to the absence of a publicly available implementation (or trained network parameters). Finally, the comparison with SOTA EBSR and BSRT shows that our DP-IR reconstructions compares favorably in terms of visual quality, while not lacking significantly in terms of fidelity.

Dynamic Scene Deblurring. Table 2 show quantitative results on the GoPro [53] and HIDE [64] datasets, respectively. We compare our approach with the SOTA reconstruction-based methods: NAFNet [59], FFTFormer [34] and diffusion-based methods: DvSR [78], InDI [15], and icDPM [59]. Our framework outperforms all competing methods across the perceptual metrics and demonstrates the best perception-distortion (P-D) trade-off among all perceptual-based methods. Moreover, our DP-IR uses twice less number of reverse steps (NFE=5) compared to the state-of-the-art InDI and still achieves better perceptual quality (e.g. LPIPS) and is more consistent with the ground-truth (+2.22dB

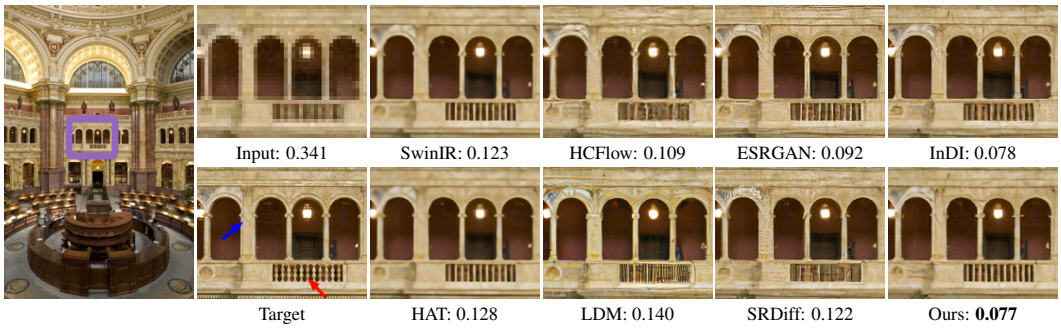

Figure 4: Visual comparisons on the DIV2K validation set for the task of $4\times$ bicubic super-resolution (best viewed by zooming in). Every output image is accompanied by its LPIPS value.

Table 3: Performance evaluation on the DIV2K validation set for $4\times$ SISR. We highlight the overall best for each metric, and the best among perceptual-oriented methods.

Methods	PSNR↑	SSIM↑	LPIPS↓	TOPIQ$_\Delta$↓	NFE ↓	Params↓
Target	∞	1	0	0	N/A	N/A
SRResNet	29.07	0.824	0.266	0.046	N/A	1.5M
RRDB	29.48	0.834	0.254	0.038	N/A	16.7M
SwinIR	29.63	0.837	0.248	0.030	N/A	11.9M
LIIF	29.30	0.830	0.258	0.046	N/A	22.3M
HAT	29.75	0.840	0.245	0.035	N/A	20.6M
Perceptual-oriented Methods						
ESRGAN	26.64	0.758	0.115	0.014	N/A	16.7M
HCFlow	27.02	0.766	0.124	0.021	N/A	23.2M
SwinIR-GAN	24.88	0.734	0.222	0.115	N/A	11.9M
LDM	23.30	0.697	0.218	0.019	100	169.0M
SRDiff	27.14	0.773	0.129	0.008	100	23.6M
InDI	26.45	0.741	0.136	0.009	100	62.3M
IDM	27.35	0.782	0.147	0.008	2000	116.6M
Ours	28.12	0.793	0.140	0.002	51	28.5M

Table 4: Ablation on various denoiser and IR networks on DIV2K validation for $4\times$ SISR.

Denoiser	IR Network	PSNR↑	SSIM↑	LPIPS↓	TOPIQ$_\Delta$↓
Target		inf	1	0	0
UDP	RRDB	27.93	0.777	0.149	0.006
MIRNet-S	RRDB	28.12	0.795	0.150	0.014
MIRNet-S	SwinIR	28.12	0.793	0.140	0.002

Table 5: Ablation on various fusion networks on DIV2K validation for $4\times$ SR task

Fusion Module	PSNR↑	SSIM↑	LPIPS↓	Params
TDWA	27.26	0.751	0.266	8385
L-DWT	29.27	0.818	0.233	30.0M
U-Net	28.19	0.785	0.141	31.0M
Proposed Fusion	27.93	0.777	0.149	0.7M

Table 6: Perception-Distortion trade-off on DIV2K validation for $4\times$ SISR.

Timestep, τ	50	100	150	200	250	300	350
PSNR ↑	28.77	28.46	28.24	28.06	27.93	27.81	27.71
LPIPS↓	0.170	0.161	0.155	0.152	0.149	0.147	0.146

in PSNR). Despite the fact that our fusion module is trained only on the GoPro dataset, the gains in perceptual quality do transfer over the HIDE test images, providing SOTA scores for the perceptual metrics. Also, among perceptual-oriented methods, DP-IR outperforms the closest competitor DvSR by 1.55dB in terms of PSNR. Visual comparisons of our method and the SOTA deblurring models: NAFNet, FFTFormer, DvSR, and InDI is depicted in Figure 3. From these results we observe that our model shows a noticeable improvement in perceptual quality. Moreover, we have performed a computational cost analysis for the diffusion-based models and significantly outperformed existing methods from $\sim$2 to 100 times (see appendix F).

Super-Resolution. We compare our method with reconstruction-based [38, 73, 42, 11, 10], GAN-based [73, 87, 42], NF-based [43] models, and DPMs [60, 39, 15, 20]. Table 3 summarizes the quantitative results on the DIV2K validation set. Our solution produces the best fidelity scores among all six perceptual-based methods and the best TOPIQ perceptual metric among all competing methods. The visual comparison in Figure 4 reveals that our framework produces super-resolved images that exhibit more refined structures and fine-grained details. Additional examples for all reported IR tasks are provided in appendix M.

5 Ablation Studies

Modular Approach. One of the main benefits of our framework is its ability to capitalize on the performance of existing restoration networks at a relatively low additional computational cost. To showcase this, we conduct an experiment on the task of $4\times$ SISR using two different denoising architectures, namely UDP [7] and MIRNet-S [81], and two IR architectures, namely RRDB [1] and SwinIR [42]. UDP is trained with the same settings as MIRNet-S and the fusion module is retrained for each of the three cases. From Table 4, we observe that the same IR network combined with a better denoising module, leads to better fidelity (see PSNR, SSIM). A same trend, but for perceptual metrics is observed if one upgrades the IR module from RRDB to the SwinIR and keeps the same denoising module. This clearly indicates that one can achieve better results by employing either a

more powerful denoiser or IR module, without the need to fully retrain the entire score estimator as is the common practice followed in most of the existing methods (*e.g.* LDM, SRDiff, DvSR, *etc.*).

Fusion Strategies. In this study we use the UDP denoiser and the RRDB network from Table 4 and study several different fusion approaches, namely the Time-dependent Weighted Averaging (TDWA), Learnable Discrete Wavelet Transform (L-DWT), our proposed Fusion network, and U-Net with time embeddings [56]. TDWA consists of sinusoidal positional encoding followed by a three-layer MLP, which predicts the weights for the timestep t. Those weights are then passed to the 3CA layer (section 3.3) together with the outputs from the denoiser and IR modules to perform the fusion in the image space. In contrast, L-DWT directly learns weights for each scale and channel of a 3-level Haar DWT [24]. L-DWT has only 30 trainable parameters, which needs significantly less training time and data. Table 5 indicates that, as expected, a more powerful fusion module leads to better perceptual and reconstruction quality. Overall, we see that the proposed Fusion network shows a better balance between the reconstruction, visual quality and the computational cost.

Perception-Distortion Trade-off. By varying the timestep τ, when the denoiser and fusion modules are activated, one can favor the perceptual quality over the reconstruction fidelity (see Table 6). Here, we use the same UDP denoiser and RRDB as the IR module and experiment on the DIV2K [1] validation for the task of $4\times$ SISR. Table 6 shows that the denoiser can operate on $\tau > 250$, which corresponds to a wider noise range than the one the denoiser is initially trained for. Furthermore, we observe the same perception-distortion trade-off for dynamic scene deblurring task (see Appendix Table 13)

Limitations Our empirical findings highlight that the optimal selection of τ is intrinsically linked to the nature of the reconstruction problem, particularly the output quality of the IR Network. While we have experimentally identified optimal parameters for each test dataset in this study, we posit that a more refined approach would involve tailoring the acceleration parameters on an individual sample basis. However, the absence of a dependable methodology for assessing the quality of the IR and Denoising Networks' outputs at specific diffusion process timesteps – especially in the absence of ground truth data – constitutes a considerable challenge. This underscores a compelling avenue for future inquiry into adaptive optimization of acceleration parameters.

Furthermore, a notable constraint of our approach is its reliance on the efficacy of the employed Denoising and IR modules. As such, for novel image restoration tasks where a pre-trained IR network is unavailable, our framework might be inapplicable. Additionally, for imaging modalities (e.g. medical imaging) lacking a trained score-matching network (denoising module), it is imperative to either fine-tune an existing module or undertake comprehensive re-training with appropriate image datasets.

6 Conclusion

We present a modular conditional diffusion probabilistic framework for IR problems along with a sampling acceleration strategy that achieves a significant speed-up during the inference stage. Our framework achieves SOTA results both quantitatively and visually on the tasks of burst JDD-SR, dynamic scene deblurring, and $4\times$ SISR without the need for re-training on a large pool of data and significant computational cost. This is mainly accomplished by utilizing pretrained models and only training a relatively small fusion module. While in this work we have not exhaustively considered all blind IR problems, we hope that our results can serve as a positive indication that the perceptual quality of the reconstructed outputs can improve significantly at the cost of only several additional NFEs, making possible a wider adoption of DPMs for IR applications even when there are tight requirements on computational complexity. Our ablation studies indicate that a variety of pretrained networks can be used with our method and further improvements on the results can be achieved by utilizing better denoising, IR, and fusion modules.

References

[1] Eirikur Agustsson and Radu Timofte. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2017.

[2] Sanjeev Arora, Rong Ge, Yingyu Liang, Tengyu Ma, and Yi Zhang. Generalization and equilibrium in generative adversarial nets (gans). In *International conference on machine learning*, pages 224–232. PMLR, 2017.

[3] Benedicte Bascle, Andrew Blake, and Andrew Zisserman. Motion deblurring and super-resolution from an image sequence. In *European conference on computer vision*, pages 571–582. Springer, 1996.

[4] Goutam Bhat, Martin Danelljan, Luc Van Gool, and Radu Timofte. Deep burst super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9209–9218, 2021.

[5] Goutam Bhat, Martin Danelljan, Fisher Yu, Luc Van Gool, and Radu Timofte. Deep reparametrization of multi-frame super-resolution and denoising. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2460–2470, 2021.

[6] Yochai Blau and Tomer Michaeli. The perception-distortion tradeoff. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6228–6237, 2018.

[7] Tim Brooks, Ben Mildenhall, Tianfan Xue, Jiawen Chen, Dillon Sharlet, and Jonathan T Barron. Unprocessing images for learned raw denoising. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11036–11045, 2019.

[8] Chaofeng Chen, Jiadi Mo, Jingwen Hou, Haoning Wu, Liang Liao, Wenxiu Sun, Qiong Yan, and Weisi Lin. Topiq: A top-down approach from semantics to distortions for image quality assessment, 2023.

[9] Honggang Chen, Xiaohai He, Linbo Qing, Yuanyuan Wu, Chao Ren, Ray E. Sheriff, and Ce Zhu. Real-world single image super-resolution: A brief review. *Information Fusion*, 79:124–145, 2022.

[10] Xiangyu Chen, Xintao Wang, Jiantao Zhou, Yu Qiao, and Chao Dong. Activating more pixels in image super-resolution transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22367–22377, 2023.

[11] Yinbo Chen, Sifei Liu, and Xiaolong Wang. Learning continuous image representation with local implicit image function. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8628–8638, 2021.

[12] Jooyoung Choi, Jungbeom Lee, Chaehun Shin, Sungwon Kim, Hyunwoo Kim, and Sungroh Yoon. Perception prioritized training of diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11472–11481, 2022.

[13] Hyungjin Chung, Byeongsu Sim, and Jong Chul Ye. Come-closer-diffuse-faster: Accelerating conditional diffusion models for inverse problems through stochastic contraction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12413–12422, 2022.

[14] Joseph Paul Cohen, Margaux Luck, and Sina Honari. Distribution matching losses can hallucinate features in medical image translation. In *International conference on medical image computing and computer-assisted intervention*, pages 529–536. Springer, 2018.

[15] Mauricio Delbracio and Peyman Milanfar. Inversion by direct iteration: An alternative to denoising diffusion for image restoration. *Transactions on Machine Learning Research*, 2023. Featured Certification.

[16] Laurent Dinh, David Krueger, and Yoshua Bengio. Nice: Non-linear independent components estimation. *arXiv preprint arXiv:1410.8516*, 2014.

[17] Akshay Dudhane, Syed Waqas Zamir, Salman Khan, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Burst image restoration and enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5759–5768, 2022.

[18] Michael Elad and Arie Feuer. Restoration of a single superresolution image from several blurred, noisy, and undersampled measured images. *IEEE transactions on image processing*, 6(12): 1646–1658, 1997.

[19] Sina Farsiu, M Dirk Robinson, Michael Elad, and Peyman Milanfar. Fast and robust multiframe super resolution. *IEEE transactions on image processing*, 13(10):1327–1344, 2004.

[20] Sicheng Gao, Xuhui Liu, Bohan Zeng, Sheng Xu, Yanjing Li, Xiaoyan Luo, Jianzhuang Liu, Xiantong Zhen, and Baochang Zhang. Implicit diffusion models for continuous super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10021–10030, 2023.

[21] Ioannis Gatopoulos, Maarten Stol, and Jakub M. Tomczak. Super-resolution variational auto-encoders, 2020.

[22] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[23] Baisong Guo, Xiaoyun Zhang, Haoning Wu, Yu Wang, Ya Zhang, and Yan-Feng Wang. Lar-sr: A local autoregressive model for image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1909–1918, 2022.

[24] Alfred Haar. *Zur theorie der orthogonalen funktionensysteme*. Georg-August-Universitat, Gottingen., 1909.

[25] Samuel W Hasinoff, Dillon Sharlet, Ryan Geiss, Andrew Adams, Jonathan T Barron, Florian Kainz, Jiawen Chen, and Marc Levoy. Burst photography for high dynamic range and low-light imaging on mobile cameras. *ACM Transactions on Graphics (ToG)*, 35(6):1–12, 2016.

[26] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, pages 6840–6851. Curran Associates, Inc., 2020.

[27] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

[28] Aapo Hyvärinen. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(24):695–709, 2005.

[29] Andrey Ignatov, Luc Van Gool, and Radu Timofte. Replacing mobile camera isp with a single deep learning model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 536–537, 2020.

[30] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 694–711. Springer, 2016.

[31] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In *Proc. NeurIPS*, 2022.

[32] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[33] Diederik P. Kingma and Max Welling. Auto-Encoding Variational Bayes. In *2nd International Conference on Learning Representations, ICLR 2014*, 2014.

[34] Lingshun Kong, Jiangxin Dong, Jianjun Ge, Mingqiang Li, and Jinshan Pan. Efficient frequency domain-based transformers for high-quality image deblurring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5886–5895, 2023.

[35] Orest Kupyn, Volodymyr Budzan, Mykola Mykhailych, Dmytro Mishkin, and Jiří Matas. Deblurgan: Blind motion deblurring using conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8183–8192, 2018.

[36] Orest Kupyn, Tetiana Martyniuk, Junru Wu, and Zhangyang Wang. Deblurgan-v2: Deblurring (orders-of-magnitude) faster and better. In *The IEEE International Conference on Computer Vision (ICCV)*, 2019.

[37] Bruno Lecouat, Jean Ponce, and Julien Mairal. Lucas-kanade reloaded: End-to-end super-resolution from raw image bursts. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2370–2379, 2021.

[38] 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.

[39] 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.

[40] Yawei Li, Yulun Zhang, Radu Timofte, Luc Van Gool, Lei Yu, Youwei Li, Xinpeng Li, Ting Jiang, Qi Wu, Mingyan Han, et al. Ntire 2023 challenge on efficient super-resolution: Methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1921–1959, 2023.

[41] Wenyi Lian and Shanglian Peng. Kernel-aware burst blind super-resolution. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4892–4902, 2023.

[42] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1833–1844, 2021.

[43] Jingyun Liang, Andreas Lugmayr, Kai Zhang, Martin Danelljan, Luc Van Gool, and Radu Timofte. Hierarchical conditional flow: A unified framework for image super-resolution and image rescaling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4076–4085, 2021.

[44] Heng Liao, Jiajin Tu, Jing Xia, Hu Liu, Xiping Zhou, Honghui Yuan, and Yuxing Hu. Ascend: a scalable and unified architecture for ubiquitous deep neural network computing : Industry track paper. In *2021 IEEE International Symposium on High-Performance Computer Architecture (HPCA)*, pages 789–801, 2021.

[45] 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.

[46] Andreas Lugmayr, Martin Danelljan, Luc Van Gool, and Radu Timofte. Srflow: Learning the super-resolution space with normalizing flow. In *European Conference on Computer Vision*, pages 715–732. Springer, 2020.

[47] Andreas Lugmayr, Martin Danelljan, and Radu Timofte. Ntire 2021 learning the super-resolution space challenge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 596–612, 2021.

[48] Calvin Luo. Understanding diffusion models: A unified perspective, 2022.

[49] Ziwei Luo, Lei Yu, Xuan Mo, Youwei Li, Lanpeng Jia, Haoqiang Fan, Jian Sun, and Shuaicheng Liu. Ebsr: Feature enhanced burst super-resolution with deformable alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 471–478, 2021.

[50] Ziwei Luo, Youwei Li, Shen Cheng, Lei Yu, Qi Wu, Zhihong Wen, Haoqiang Fan, Jian Sun, and Shuaicheng Liu. Bsrt: Improving burst super-resolution with swin transformer and flow-guided deformable alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 998–1008, 2022.

[51] Nancy Mehta, Akshay Dudhane, Subrahmanyam Murala, Syed Waqas Zamir, Salman Khan, and Fahad Shahbaz Khan. Adaptive feature consolidation network for burst super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1279–1286, 2022.

[52] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. SDEdit: Guided image synthesis and editing with stochastic differential equations. In *International Conference on Learning Representations*, 2022.

[53] Seungjun Nah, Tae Hyun Kim, and Kyoung Mu Lee. Deep multi-scale convolutional neural network for dynamic scene deblurring. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[54] A. Nguyen, J. Clune, Y. Bengio, A. Dosovitskiy, and J. Yosinski. Plug & play generative networks: Conditional iterative generation of images in latent space. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3510–3520, Los Alamitos, CA, USA, 2017. IEEE Computer Society.

[55] Alex Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. *arXiv preprint arXiv:2102.09672*, 2021.

[56] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8162–8171. PMLR, 2021.

[57] Mangal Prakash, Alexander Krull, and Florian Jug. Fully unsupervised diversity denoising with convolutional variational autoencoders. In *International Conference on Learning Representations*, 2020.

[58] Dongwei Ren, Kai Zhang, Qilong Wang, Qinghua Hu, and Wangmeng Zuo. Neural blind deconvolution using deep priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3341–3350, 2020.

[59] Mengwei Ren, Mauricio Delbracio, Hossein Talebi, Guido Gerig, and Peyman Milanfar. Multiscale structure guided diffusion for image deblurring. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10721–10733, 2023.

[60] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, 2022.

[61] 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.

[62] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *Advances in neural information processing systems*, 29, 2016.

[63] Richard R Schultz and Robert L Stevenson. Extraction of high-resolution frames from video sequences. *IEEE transactions on image processing*, 5(6):996–1011, 1996.

[64] Ziyi Shen, Wenguan Wang, Jianbing Shen, Haibin Ling, Tingfa Xu, and Ling Shao. Human-aware motion deblurring. In *IEEE International Conference on Computer Vision*, 2019.

[65] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *Proceedings of the 32nd International Conference on Machine Learning*, pages 2256–2265. PMLR, 2015.

[66] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021.

[67] 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.

[68] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8934–8943, 2018.

[69] Fu-Jen Tsai, Yan-Tsung Peng, Yen-Yu Lin, Chung-Chi Tsai, and Chia-Wen Lin. Stripformer: Strip transformer for fast image deblurring. In *European Conference on Computer Vision*, pages 146–162. Springer, 2022.

[70] R Tsai. Multiframe image restoration and registration. *Advance Computer Visual and Image Processing*, 1:317–339, 1984.

[71] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[72] Pascal Vincent. A connection between score matching and denoising autoencoders. *Neural Computation*, 23(7):1661–1674, 2011.

[73] 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.

[74] Xintao Wang, Kelvin CK Chan, Ke Yu, Chao Dong, and Chen Change Loy. Edvr: Video restoration with enhanced deformable convolutional networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019.

[75] Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1905–1914, 2021.

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

[77] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4): 600–612, 2004.

[78] Jay Whang, Mauricio Delbracio, Hossein Talebi, Chitwan Saharia, Alexandros G Dimakis, and Peyman Milanfar. Deblurring via stochastic refinement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16293–16303, 2022.

[79] Bartlomiej Wronski, Ignacio Garcia-Dorado, Manfred Ernst, Damien Kelly, Michael Krainin, Chia-Kai Liang, Marc Levoy, and Peyman Milanfar. Handheld multi-frame super-resolution. *ACM Transactions on Graphics (TOG)*, 38(4):1–18, 2019.

[80] Bin Xia, Yulun Zhang, Shiyin Wang, Yitong Wang, Xinglong Wu, Yapeng Tian, Wenming Yang, and Luc Van Gool. Diffir: Efficient diffusion model for image restoration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 13095–13105, 2023.

[81] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. Learning enriched features for real image restoration and enhancement. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXV 16*, pages 492–511. Springer, 2020.

[82] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. Multi-stage progressive image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14821–14831, 2021.

[83] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5728–5739, 2022.

[84] Lujun Zhai, Yonghui Wang, Suxia Cui, and Yu Zhou. A comprehensive review of deep learning-based real-world image restoration. *IEEE Access*, 11:21049–21067, 2023.

[85] Kaihao Zhang, Wenhan Luo, Yiran Zhong, Lin Ma, Bjorn Stenger, Wei Liu, and Hongdong Li. Deblurring by realistic blurring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2737–2746, 2020.

[86] 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 (CVPR)*, 2018.

[87] Wenlong Zhang, Yihao Liu, Chao Dong, and Yu Qiao. Ranksrgan: Super resolution generative adversarial networks with learning to rank. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):7149–7166, 2021.

[88] Hang Zhao, Orazio Gallo, Iuri Frosio, and Jan Kautz. Loss functions for image restoration with neural networks. *IEEE Transactions on computational imaging*, 3(1):47–57, 2016.

[89] Zhengxia Zou, Keyan Chen, Zhenwei Shi, Yuhong Guo, and Jieping Ye. Object detection in 20 years: A survey. *Proceedings of the IEEE*, 111(3):257–276, 2023.

A Conditional DPM Training Requirements

To the best of our knowledge, all diffusion-based approaches that have been proposed in the literature to deal with image restoration tasks, require the training of far larger conditional backbone networks ($\sim$10-100M params). This turns out to be significantly more challenging both in terms of necessary training data and computational resources. To showcase this, we provide an indicative example below. If we adopt the existing diffusion-based SISR baselines and train them for a completely different restoration problem, by following the original authors' training strategies it turns out that the computational and data requirements are significantly higher than those of our method.

Table 7: Comparison of the proposed approach against existing DPM methods for SISR task in terms of training dataset requirements and training parameters.

Method	Params required	Data required
Ours	1x	1x
SRDiff	$\sim$ 34x	$\sim$ 4x
LDM	$\sim$ 240x	$\sim$ 1000x
InDI	$\sim$ 89x	$\sim$ 1x
IDM	$\sim$ 167x	$\sim$ 1x

Based on these data, we can safely state that our strategy provides a reasonable trade-off between the required training complexity and the competitive performance of our method to a variety of blind inverse problems.

B Proof of Lemma 3.1

To derive the conditional score function, we first express the conditional probability $p\left(\boldsymbol{x}_t|\boldsymbol{y}\right)$ as:

$$p\left(\boldsymbol{x}_t|\boldsymbol{y}\right) = \int p\left(\boldsymbol{x}_t, \boldsymbol{x}_0|\boldsymbol{y}\right)\, d\boldsymbol{x}_0 = \int p\left(\boldsymbol{x}_t|\boldsymbol{y}, \boldsymbol{x}_0\right) p\left(\boldsymbol{x}_0|\boldsymbol{y}\right)\, d\boldsymbol{x}_0 = \int q\left(\boldsymbol{x}_t|\boldsymbol{x}_0\right) p\left(\boldsymbol{x}_0|\boldsymbol{y}\right)\, d\boldsymbol{x}_0, \tag{12}$$

where we used the fact that $\boldsymbol{x}_t$ is conditionally independent of $\boldsymbol{y}$ and according to eq. (2) it holds $p\left(\boldsymbol{x}_t|\boldsymbol{y}, \boldsymbol{x}_0\right) = q\left(\boldsymbol{x}_t|\boldsymbol{x}_0\right)$. Differentiating both sides of eq. (12) with respect to (w.r.t.) $\boldsymbol{x}_t$ we get:

$$\nabla_{\boldsymbol{x}_t} p\left(\boldsymbol{x}_t|\boldsymbol{y}\right) = \int p\left(\boldsymbol{x}_0|\boldsymbol{y}\right) \nabla_{\boldsymbol{x}_t} q\left(\boldsymbol{x}_t|\boldsymbol{x}_0\right)\, d\boldsymbol{x}_0. \tag{13}$$

Based on the definition of $q\left(\boldsymbol{x}_t|\boldsymbol{x}_0\right)$ in eq. (2), it also holds that: $\nabla_{\boldsymbol{x}_t} q\left(\boldsymbol{x}_t|\boldsymbol{x}_0\right) = -q\left(\boldsymbol{x}_t|\boldsymbol{x}_0\right) \frac{\boldsymbol{x}_t - \sqrt{\bar{\alpha}_t}\boldsymbol{x}_0}{1-\bar{\alpha}_t}$. Substituting this result back to eq. (13), we get:

$$\nabla_{\boldsymbol{x}_t} p\left(\boldsymbol{x}_t|\boldsymbol{y}\right) = \frac{\sqrt{\bar{\alpha}_t}}{1-\bar{\alpha}_t} \int \boldsymbol{x}_0 q\left(\boldsymbol{x}_t|\boldsymbol{x}_0\right) p\left(\boldsymbol{x}_0|\boldsymbol{y}\right)\, d\boldsymbol{x}_0 - \frac{\boldsymbol{x}_t}{1-\bar{\alpha}_t} \overbrace{\int q\left(\boldsymbol{x}_t|\boldsymbol{x}_0\right) p\left(\boldsymbol{x}_0|\boldsymbol{y}\right)\, d\boldsymbol{x}_0}^{p(\boldsymbol{x}_t|\boldsymbol{y})}. \tag{14}$$

Next, if we divide both sides in eq. (13) with $p\left(\boldsymbol{x}_t|\boldsymbol{y}\right)$ and use that $\nabla_{\boldsymbol{x}_t} \log p\left(\boldsymbol{x}_t|\boldsymbol{y}\right) = \frac{\nabla_{\boldsymbol{x}_t} p(\boldsymbol{x}_t|\boldsymbol{y})}{p(\boldsymbol{x}_t|\boldsymbol{y})}$, we get:

$$\nabla_{\boldsymbol{x}_t} \log p\left(\boldsymbol{x}_t|\boldsymbol{y}\right) = \frac{1}{\left(1-\bar{\alpha}_t\right)} \left(\sqrt{\bar{\alpha}_t} \int \boldsymbol{x}_0 \frac{q\left(\boldsymbol{x}_t|\boldsymbol{x}_0\right) p\left(\boldsymbol{x}_0|\boldsymbol{y}\right)}{p\left(\boldsymbol{x}_t|\boldsymbol{y}\right)}\, d\boldsymbol{x}_0 - \boldsymbol{x}_t \right). \tag{15}$$

We can further express the integral in eq. (15) as follows:

$$\int \boldsymbol{x}_0 \frac{q\left(\boldsymbol{x}_t|\boldsymbol{x}_0\right) p\left(\boldsymbol{x}_0|\boldsymbol{y}\right)}{p\left(\boldsymbol{x}_t|\boldsymbol{y}\right)}\, d\boldsymbol{x}_0 = \int \boldsymbol{x}_0 \frac{q\left(\boldsymbol{x}_t|\boldsymbol{x}_0\right) p\left(\boldsymbol{x}_0|\boldsymbol{y}\right) p\left(\boldsymbol{y}\right)}{p\left(\boldsymbol{x}_t|\boldsymbol{y}\right) p\left(\boldsymbol{y}\right)}\, d\boldsymbol{x}_0 = \int \boldsymbol{x}_0 \frac{p\left(\boldsymbol{x}_t|\boldsymbol{x}_0, \boldsymbol{y}\right) p\left(\boldsymbol{x}_0, \boldsymbol{y}\right)}{p\left(\boldsymbol{x}_t, \boldsymbol{y}\right)}\, d\boldsymbol{x}_0$$
$$= \int \boldsymbol{x}_0 \frac{p\left(\boldsymbol{x}_t, \boldsymbol{x}_0, \boldsymbol{y}\right)}{p\left(\boldsymbol{x}_t, \boldsymbol{y}\right)}\, d\boldsymbol{x}_0 = \int \boldsymbol{x}_0 p\left(\boldsymbol{x}_0|\boldsymbol{y}, \boldsymbol{x}_t\right)\, d\boldsymbol{x}_0 = \mathbb{E}\left[\boldsymbol{x}_0|\boldsymbol{y}, \boldsymbol{x}_t\right].$$

Substituting this result in eq. (15) finally leads us to the result of the lemma: $\nabla_{\boldsymbol{x}_t} p\left(\boldsymbol{x}_t|\boldsymbol{y}\right) = \frac{\sqrt{\bar{\alpha}_t}\, \mathbb{E}[\boldsymbol{x}_0|\boldsymbol{y}, \boldsymbol{x}_t] - \boldsymbol{x}_t}{1-\bar{\alpha}_t}$.

C Theoretical Justification of the Conditional Expectation Approximation

By construction, for the forward process from eq. (1) it holds that $\boldsymbol{x}_T \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_N)$ when $T \to \infty$, which means that in the beginning of reverse sampling, the latent variable $\boldsymbol{x}_T$ does not contain any information about $\boldsymbol{x}_0$. Below we provide a theoretical result that can serve as an indication that during the sampling process and up to some timestep τ, we can use an approximation of $\mathbb{E}[\boldsymbol{x}_0|\boldsymbol{y}, \boldsymbol{x}_t] \approx \mathbb{E}[\boldsymbol{x}_0|\boldsymbol{y}]$, given that the contribution of $\mathbb{E}[\boldsymbol{x}_0|\boldsymbol{x}_t]$ is not significant enough. This is achieved by disabling the lower branch of our proposed conditional score matching network, which includes the Denoising $\boldsymbol{\phi}^D_{\boldsymbol{\theta}_D}(\tilde{\boldsymbol{x}}_t, \tilde{\sigma}_t)$ and Fusion $\boldsymbol{\phi}^F_{\boldsymbol{\theta}_F}\left(\boldsymbol{x}^{IR}_0, \boldsymbol{x}^D_0\right)$ modules (Figure 1). A formal theoretical analysis is available under the following simplifications:

- The observation model is linear: $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}_0 + \boldsymbol{n}$.
- Denoising Network is an identity transformation: $\boldsymbol{\phi}^D_{\boldsymbol{\theta}_D}(\tilde{\boldsymbol{x}}_t, \tilde{\sigma}_t) \equiv \tilde{\boldsymbol{x}}_t$.
- Image Restoration Network is a back-projection: $\boldsymbol{\phi}^{IR}_{\boldsymbol{\theta}_{IR}}(\boldsymbol{y}) \equiv \boldsymbol{A}^T\boldsymbol{y}$
- Fusion Network is a convex combination in a spatial domain: $\boldsymbol{\phi}^F_{\boldsymbol{\theta}_F}\left(\boldsymbol{x}^{IR}_0, \boldsymbol{x}^D_0, t\right) \equiv w\boldsymbol{x}^{IR}_0 + (1-w)\boldsymbol{x}^D_0$.
- Diffusion process approaches the continuous-time regime: $T \to \infty$.

Our theoretical result is provided in the form of the following proposition.

Proposition C.1. *Let $\boldsymbol{y} \in \mathbb{R}^M$ be the measurements obtained according to the following observation model: $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x_0} + \boldsymbol{n}$, where $\boldsymbol{A} \in \mathbb{R}^{M\times N}$, $\|\boldsymbol{A}\|_2 \leq 1$ and $\boldsymbol{n} \sim \mathcal{N}(\mathbf{0}, \sigma_{\boldsymbol{y}}\boldsymbol{I}_M)$, $\sigma_{\boldsymbol{y}} \leq 1$. Then, for any given $\boldsymbol{A}$,$\sigma_{\boldsymbol{y}}$,$\boldsymbol{x}_0 \in \mathbb{R}^N$, and any fixed $w \in (0,1)$, there exists a timestep τ, such that for all $t > \tau$ the back-projected signal $\boldsymbol{A}^T\boldsymbol{y}$ approximates $\boldsymbol{x}_0$ better than the convex combination $w\boldsymbol{A}^T\boldsymbol{y} + (1-w)\tilde{\boldsymbol{x}}_t$ in the following sense:*

$$\mathbb{E}_{\boldsymbol{y}|\boldsymbol{x}_0}\|\boldsymbol{x}_0 - \boldsymbol{A}^T\boldsymbol{y}\|_2^2 \leq \mathbb{E}_{\boldsymbol{y},\boldsymbol{x}_t|\boldsymbol{x}_0}\|\boldsymbol{x}_0 - \left(w\boldsymbol{A}^T\boldsymbol{y} + (1-w)\tilde{\boldsymbol{x}}_t\right)\|_2^2, \tag{16}$$

where $\tilde{\boldsymbol{x}}_t \sim \mathcal{N}\left(\boldsymbol{x}_0, \tilde{\sigma}_t^2\boldsymbol{I}_N\right), \tilde{\sigma}_t^2 \equiv \frac{1-\bar{\alpha}_t}{\bar{\alpha}_t}$ is defined as a noisy version of $\boldsymbol{x}_0$ within the diffusion process described by eqs. (1) *and* (2) *with $T \to \infty$.*

For our proof we use the following intermediate result.

Lemma C.2. *Let $\boldsymbol{\sigma} \in \mathbb{R}^n_+$, $\boldsymbol{x} \sim p(\boldsymbol{x}) \equiv \mathcal{N}\left(\mathbf{0}, \operatorname{diag}\left(\boldsymbol{\sigma}^2\right)\right)$, $\boldsymbol{B} \in \mathbb{R}^{n\times n}$. Then the following holds:*

$$\int \boldsymbol{x}^\mathsf{T}\boldsymbol{B}\boldsymbol{x}\, p(\boldsymbol{x})\, d\boldsymbol{x} = \operatorname{tr}\left(\operatorname{diag}\left(\boldsymbol{\sigma}^2\right)\boldsymbol{B}\right). \tag{17}$$

Proof. We first note that since $\operatorname{Cov}(\boldsymbol{x}) = \operatorname{diag}\left(\boldsymbol{\sigma}^2\right)$, the random variables $\boldsymbol{x}_i$ are independent, $p(\boldsymbol{x}) = \prod_{i=1}^n p(\boldsymbol{x}_i)$, and distributed as $\boldsymbol{x}_i \sim p(\boldsymbol{x}_i) \equiv \mathcal{N}\left(0, \boldsymbol{\sigma}_i^2\right)$. We further note that it holds:

$$\boldsymbol{x}^\mathsf{T}\boldsymbol{B}\boldsymbol{x} = \sum_{i,j=1}^n \boldsymbol{x}_i\boldsymbol{x}_j\boldsymbol{B}_i^j = \sum_{\substack{i,j=1\\ i\neq j}}^n \boldsymbol{x}_i\boldsymbol{x}_j\boldsymbol{B}_i^j + \sum_{i=1}^n \boldsymbol{x}_i^2\boldsymbol{B}_i^i. \tag{18}$$

Using these results, the derivation of eq. (17) is straightforward:

$$\int \boldsymbol{x}^\mathsf{T}\boldsymbol{B}\boldsymbol{x}\, p(\boldsymbol{x})\, d\boldsymbol{x} = \sum_{\substack{i,j=1\\ i\neq j}}^n \left(\boldsymbol{B}_i^j \int \boldsymbol{x}_i p(\boldsymbol{x}_i)\, d\boldsymbol{x}_i \int \boldsymbol{x}_j p(\boldsymbol{x}_j)\, d\boldsymbol{x}_j \prod_{\substack{k=1\\ k\neq i,j}}^n \int p(\boldsymbol{x}_k)\, d\boldsymbol{x}_k\right) \tag{19}$$

$$+ \sum_{i=1}^n \boldsymbol{B}_i^i \int \boldsymbol{x}_i^2 p(\boldsymbol{x}_i)\, d\boldsymbol{x}_i \prod_{\substack{k=1\\ k\neq i}}^n \int p(\boldsymbol{x}_k)\, d\boldsymbol{x}_k = \sum_{i=1}^n \boldsymbol{B}_i^i\boldsymbol{\sigma}_i^2 = \operatorname{tr}\left(\operatorname{diag}\left(\boldsymbol{\sigma}^2\right)\boldsymbol{B}\right),$$

since for all $i = \overline{1,n}$, it holds that $\int \boldsymbol{p}(\boldsymbol{x}_i)\, d\boldsymbol{x}_i = 1$, $\int \boldsymbol{x}_i p(\boldsymbol{x}_i)\, d\boldsymbol{x}_i = 0$, $\int \boldsymbol{x}_i^2 p(\boldsymbol{x}_i)\, d\boldsymbol{x}_i = \boldsymbol{\sigma}_i^2$. □

We start our proof from computing the left part of the inequality in eq. (16). Since $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}_0 + \boldsymbol{n}$, where $\boldsymbol{n} \sim \mathcal{N}(\mathbf{0}, \sigma_{\boldsymbol{y}}\boldsymbol{I}_M)$, we have that: $p(\boldsymbol{y}|\boldsymbol{x}_0) = \mathcal{N}\left(\boldsymbol{A}\boldsymbol{x}_0, \sigma_{\boldsymbol{y}}^2\boldsymbol{I}_M\right)$. This allows us to write:

$$\mathbb{E}_{\boldsymbol{y}|\boldsymbol{x}_0}\|\boldsymbol{x}_0 - \boldsymbol{A}^\mathsf{T}\boldsymbol{y}\|_2^2 = \int \|\boldsymbol{x}_0 - \boldsymbol{A}^\mathsf{T}\boldsymbol{y}\|_2^2\, p(\boldsymbol{y}|\boldsymbol{x}_0)\, d\boldsymbol{y} = \int \|\boldsymbol{x}_0 - \boldsymbol{A}^\mathsf{T}\boldsymbol{y}\|_2^2\, \frac{\exp\left(-\frac{\|\boldsymbol{y}-\boldsymbol{A}\boldsymbol{x}_0\|_2^2}{2\sigma_{\boldsymbol{y}}^2}\right)}{\left(\sqrt{2\pi}\sigma_{\boldsymbol{y}}\right)^M}\, d\boldsymbol{y} \equiv \mathcal{I}_1.$$

We now apply a change of variables from $\boldsymbol{y}$ to $\boldsymbol{n} = \boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}_0$, for which $d\boldsymbol{n} = d\boldsymbol{y}$. Using the fact that $\frac{1}{\left(\sqrt{2\pi}\sigma_{\boldsymbol{y}}\right)^M}\exp\left(-\frac{\|\boldsymbol{n}\|_2^2}{2\sigma_{\boldsymbol{y}}^2}\right) = p\left(\boldsymbol{n}\right) = \mathcal{N}\left(\boldsymbol{0}, \sigma_{\boldsymbol{y}}^2\boldsymbol{I}_M\right)$, we have that:

$$\mathcal{I}_1 = \int \|\boldsymbol{x}_0 - \boldsymbol{A}^\mathsf{T}\boldsymbol{A}\boldsymbol{x}_0 - \boldsymbol{A}^\mathsf{T}\boldsymbol{n}\|_2^2 \, p\left(\boldsymbol{n}\right) \, d\boldsymbol{n}. \tag{20}$$

Next, we denote $\Delta\boldsymbol{x}_0 \equiv \boldsymbol{x}_0 - \boldsymbol{A}^\mathsf{T}\boldsymbol{A}\boldsymbol{x}_0 = \left(\boldsymbol{I}_N - \boldsymbol{A}^\mathsf{T}\boldsymbol{A}\right)\boldsymbol{x}_0$ and expand the norm inside the integral to get:

$$\mathcal{I}_1 = \Delta\boldsymbol{x}_0^\mathsf{T}\Delta\boldsymbol{x}_0 \int p\left(\boldsymbol{n}\right) \, d\boldsymbol{n} - 2\Delta\boldsymbol{x}_0^\mathsf{T}\boldsymbol{A}^\mathsf{T}\int \boldsymbol{n}p\left(\boldsymbol{n}\right) \, d\boldsymbol{n} + \int \boldsymbol{n}^\mathsf{T}\boldsymbol{A}\boldsymbol{A}^\mathsf{T}\boldsymbol{n}p\left(\boldsymbol{n}\right) \, d\boldsymbol{n}.$$

Since $p\left(\boldsymbol{n}\right) = \mathcal{N}\left(\boldsymbol{0}, \sigma_{\boldsymbol{y}}^2\boldsymbol{I}_M\right)$ is a zero-mean Gaussian probability distribution, it holds that $\int p\left(\boldsymbol{n}\right) \, d\boldsymbol{n} = 1, \int \boldsymbol{n}p\left(\boldsymbol{n}\right) \, d\boldsymbol{n} = 0$, and we get:

$$\mathcal{I}_1 = \Delta\boldsymbol{x}_0^\mathsf{T}\Delta\boldsymbol{x}_0 + \int \boldsymbol{n}^\mathsf{T}\boldsymbol{A}\boldsymbol{A}^\mathsf{T}\boldsymbol{n}p\left(\boldsymbol{n}\right) \, d\boldsymbol{n} = \Delta\boldsymbol{x}_0^\mathsf{T}\Delta\boldsymbol{x}_0 + \mathrm{tr}\left(\sigma_{\boldsymbol{y}}^2\boldsymbol{A}\boldsymbol{A}^\mathsf{T}\right) = \Delta\boldsymbol{x}_0^\mathsf{T}\Delta\boldsymbol{x}_0 + \sigma_{\boldsymbol{y}}^2\|\boldsymbol{A}\|_F^2, \tag{21}$$

where we have used the result of lemma C.2. Substituting back the value of $\Delta\boldsymbol{x}_0$, we end up with:

$$\mathbb{E}_{\boldsymbol{y}|\boldsymbol{x}_0}\|\boldsymbol{x}_0 - \boldsymbol{A}^\mathsf{T}\boldsymbol{y}\|_2^2 = \|\left(\boldsymbol{I}_N - \boldsymbol{A}^\mathsf{T}\boldsymbol{A}\right)\boldsymbol{x}_0\|_2^2 + \sigma_{\boldsymbol{y}}^2\|\boldsymbol{A}\|_F^2. \tag{22}$$

Next, we compute the right part of the inequality in eq. (16). We first note, that the quantities $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}_0 + \boldsymbol{n}$ and $\boldsymbol{x}_t = \sqrt{\bar{\alpha}_t}\boldsymbol{x}_0 + \boldsymbol{\epsilon}_t$ are independent when conditioned on $\boldsymbol{x}_0$, as $\boldsymbol{n} \sim \mathcal{N}\left(\boldsymbol{0}, \sigma_{\boldsymbol{y}}^2\boldsymbol{I}_M\right)$ and $\boldsymbol{\epsilon}_t \sim \mathcal{N}\left(\boldsymbol{0}, \left(1-\bar{\alpha}_t\right)\boldsymbol{I}_N\right)$ are independent random noise vectors. This allows us to write: $p\left(\boldsymbol{y}, \boldsymbol{x}_t|\boldsymbol{x}_0\right) = p\left(\boldsymbol{y}|\boldsymbol{x}_0\right)p\left(\boldsymbol{x}_t|\boldsymbol{x}_0\right)$. Using these results, the expectation $\mathbb{E}_{\boldsymbol{y},\boldsymbol{x}_t|\boldsymbol{x}_0}\|\boldsymbol{x}_0 - \left(w\boldsymbol{A}^\mathsf{T}\boldsymbol{y} + (1-w)\tilde{\boldsymbol{x}}_t\right)\|_2^2$ can be written as:

$$\begin{aligned}&\mathbb{E}_{\boldsymbol{y},\boldsymbol{x}_t|\boldsymbol{x}_0}\|\boldsymbol{x}_0 - \left(w\boldsymbol{A}^\mathsf{T}\boldsymbol{y} + (1-w)\tilde{\boldsymbol{x}}_t\right)\|_2^2 \\ =&\mathcal{I}_2 \equiv \int \|\boldsymbol{x}_0 - \left(w\boldsymbol{A}^\mathsf{T}\boldsymbol{y} + (1-w)\frac{\boldsymbol{x}_t}{\sqrt{\bar{\alpha}_t}}\right)\|_2^2 \, p\left(\boldsymbol{y}|\boldsymbol{x}_0\right)p\left(\boldsymbol{x}_t|\boldsymbol{x}_0\right) \, d\boldsymbol{y}d\boldsymbol{x}_t.\end{aligned} \tag{23}$$

Similarly to eq. (20), we apply the following change of integration variables: $\boldsymbol{n} = \boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}_0$, $d\boldsymbol{y} = d\boldsymbol{n}$ and $\boldsymbol{\epsilon}_t = \boldsymbol{x}_t - \sqrt{\bar{\alpha}_t}\boldsymbol{x}_0$, $d\boldsymbol{x}_t = d\boldsymbol{\epsilon}_t$. We additionally note, that $\boldsymbol{x}_0 - \left(w\boldsymbol{A}^\mathsf{T}\boldsymbol{y} + \frac{(1-w)}{\sqrt{\bar{\alpha}_t}}\boldsymbol{x}_t\right) = \boldsymbol{x}_0 - w\boldsymbol{A}^\mathsf{T}\boldsymbol{A}\boldsymbol{x}_0 - (1-w)\,\boldsymbol{x}_0 - w\boldsymbol{A}^\mathsf{T}\boldsymbol{n} - \frac{1-w}{\sqrt{\bar{\alpha}_t}}\boldsymbol{\epsilon}_t$, and that $\boldsymbol{x}_0 - w\boldsymbol{A}^\mathsf{T}\boldsymbol{A}\boldsymbol{x}_0 - (1-w)\,\boldsymbol{x}_0 = w\left(\boldsymbol{I}_N - \boldsymbol{A}^\mathsf{T}\boldsymbol{A}\right)\boldsymbol{x}_0 = w\Delta\boldsymbol{x}_0$. As a result, the integral $\mathcal{I}_2$ takes the form:

$$\mathcal{I}_2 = \int \|w\Delta\boldsymbol{x}_0 - w\boldsymbol{A}^\mathsf{T}\boldsymbol{n} - \frac{1-w}{\sqrt{\bar{\alpha}_t}}\boldsymbol{\epsilon}_t\|_2^2 \, p\left(\boldsymbol{n}\right)p\left(\boldsymbol{\epsilon}_t\right) \, d\boldsymbol{n}d\boldsymbol{\epsilon}_t. \tag{24}$$

Now we use the augmented variable

$$\boldsymbol{\epsilon} = \begin{bmatrix}\boldsymbol{n}^\mathsf{T} & \boldsymbol{\epsilon}_t^\mathsf{T}\end{bmatrix}^\mathsf{T} \sim p\left(\boldsymbol{n}\right)p\left(\boldsymbol{\epsilon}_t\right) = \mathcal{N}\left(\boldsymbol{0}, \mathrm{diag}\left(\begin{bmatrix}\sigma_{\boldsymbol{y}}^2\boldsymbol{1}_M^\mathsf{T} & \left(1-\bar{\alpha}_t\right)\boldsymbol{1}_N^\mathsf{T}\end{bmatrix}^\mathsf{T}\right)\right) = p\left(\boldsymbol{\epsilon}\right),$$

where with $\boldsymbol{1}_i$ we denote the vector of dimension i filled with ones. We also denote with $\boldsymbol{W} = \begin{bmatrix}w\boldsymbol{A}^T & \frac{1-w}{\sqrt{\bar{\alpha}_t}}\boldsymbol{I}_N\end{bmatrix}$, as in this case, $\boldsymbol{W}\boldsymbol{\epsilon} = w\boldsymbol{A}^\mathsf{T}\boldsymbol{n} + \frac{1-w}{\sqrt{\bar{\alpha}_t}}\boldsymbol{\epsilon}_t$. With all these modifications, the integral $\mathcal{I}_2$ takes the form:

$$\begin{aligned}\mathcal{I}_2 = &\int \|w\Delta\boldsymbol{x}_0 - \boldsymbol{W}\boldsymbol{\epsilon}\|_2^2 \, p\left(\boldsymbol{\epsilon}\right) \, d\boldsymbol{\epsilon} = w^2\Delta\boldsymbol{x}_0^\mathsf{T}\boldsymbol{x}_0 \int p\left(\boldsymbol{\epsilon}\right) \, d\boldsymbol{\epsilon} - 2w\Delta\boldsymbol{x}_0^\mathsf{T}\boldsymbol{W}\int \boldsymbol{\epsilon}\, p\left(\boldsymbol{\epsilon}\right) \, d\boldsymbol{\epsilon} \\ &+ \int \boldsymbol{\epsilon}^\mathsf{T}\boldsymbol{W}^\mathsf{T}\boldsymbol{W}\boldsymbol{\epsilon}\, p\left(\epsilon\right) \, d\boldsymbol{\epsilon} = w^2\boldsymbol{\Delta}\boldsymbol{x}_0^\mathsf{T}\Delta\boldsymbol{x}_0 + \mathrm{tr}\left(\mathrm{diag}\left(\begin{bmatrix}\sigma_{\boldsymbol{y}}^2\boldsymbol{1}_M^\mathsf{T} & \left(1-\bar{\alpha}_t\right)\boldsymbol{1}_N^\mathsf{T}\end{bmatrix}^\mathsf{T}\right)\boldsymbol{W}^\mathsf{T}\boldsymbol{W}\right).\end{aligned} \tag{25}$$

Further, we define $\boldsymbol{D} = \mathrm{diag}\left(\begin{bmatrix}\sigma_{\boldsymbol{y}}^2\boldsymbol{1}_M^\mathsf{T} & \left(1-\bar{\alpha}_t\right)\boldsymbol{1}_N^\mathsf{T}\end{bmatrix}^\mathsf{T}\right)$ and compute the respective trace in eq. (25) as:

$$\begin{aligned}&\mathrm{tr}\left(\boldsymbol{D}\boldsymbol{W}^\mathsf{T}\boldsymbol{W}\right) = \mathrm{tr}\left(\boldsymbol{W}\boldsymbol{D}\boldsymbol{W}^\mathsf{T}\right)\mathrm{tr}\left(\left(\boldsymbol{W}\boldsymbol{D}^{1/2}\right)\left(\boldsymbol{W}\boldsymbol{D}^{1/2}\right)^\mathsf{T}\right)\|\boldsymbol{W}\boldsymbol{D}^{1/2}\|_F^2 \\ &= w^2\sigma_{\boldsymbol{y}}^2\|\boldsymbol{A}\|_F^2 + (1-w)^2\frac{1-\bar{\alpha}_t}{\bar{\alpha}_t}\|\boldsymbol{I}_N\|_F^2 = w^2\sigma_{\boldsymbol{y}}^2\|\boldsymbol{A}\|_F^2 + (1-w)^2\frac{1-\bar{\alpha}_t}{\bar{\alpha}_t}N.\end{aligned}$$

Based on the above, the integral $\mathcal{I}_2$ takes the form:

$$\mathcal{I}_2 = w^2 \Delta \boldsymbol{x}_0^\mathsf{T} \Delta \boldsymbol{x}_0 + w^2 \sigma_{\boldsymbol{y}}^2 \|\boldsymbol{A}\|_F^2 + (1-w)^2 \frac{1-\bar{\alpha}_t}{\bar{\alpha}_t} = w^2 \mathcal{I}_1 + \left(1-w^2\right) \frac{1-\bar{\alpha}_t}{\bar{\alpha}_t} N. \tag{26}$$

As a result of our derivations, the inequality from Proposition C.1 takes the form:

$$\mathbb{E}_{\boldsymbol{y}|\boldsymbol{x}_0} \|\boldsymbol{x}_0 - \boldsymbol{A}^\mathsf{T} \boldsymbol{y}\|_2^2 \leq w^2 \, \mathbb{E}_{\boldsymbol{y}|\boldsymbol{x}_0} \|\boldsymbol{x}_0 - \boldsymbol{A}^\mathsf{T} \boldsymbol{y}\|_2^2 + (1-w)^2 \frac{1-\bar{\alpha}_t}{\bar{\alpha}_t} N. \tag{27}$$

The quantity $\mathbb{E}_{\boldsymbol{y}|\boldsymbol{x}_0} \|\boldsymbol{x}_0 - \boldsymbol{A}^\mathsf{T} \boldsymbol{y}\|_2^2$ has finite value, as it can be upper bounded as follows:

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{y}|\boldsymbol{x}_0} \|\boldsymbol{x}_0 - \boldsymbol{A}^\mathsf{T} \boldsymbol{y}\|_2^2 = \|\left(\boldsymbol{I}_N - \boldsymbol{A}^\mathsf{T} \boldsymbol{A}\right) \boldsymbol{x}_0\|_2^2 + \sigma_{\boldsymbol{y}}^2 \|\boldsymbol{A}\|_F^2 \\
\leq &\|\boldsymbol{I}_N - \boldsymbol{A}^\mathsf{T} \boldsymbol{A}\|_2^2 \|\boldsymbol{x}_0\|_2^2 + \sigma_{\boldsymbol{y}}^2 \min\left(M, N\right) \|\boldsymbol{A}\|_2^2 \\
\leq &\left(\|\boldsymbol{I}_N\|_2 + \|\boldsymbol{A}\|_2^2\right)^2 N + \min\left(M, N\right) \|\boldsymbol{A}\|_2^2 \leq 4N + \min\left(M, N\right).
\end{aligned}$$

In the above chain of inequalities we have used the submultiplicative property and triangle inequality for ℓ_2 matrix norms together with the norms equivalence inequality for ℓ_2 and Frobenius matrix norms, where we also upper-bound the rank of matrix $\boldsymbol{A}$ with its minimal dimension. Additionally, we have used the assumptions of Proposition C.1, specifically $\sigma_{\boldsymbol{y}} \leq 1$ and $\|\boldsymbol{A}\| \leq 1$.

As to the second term of the rhs of eq. (27), we note that since $\lim_{t \to \infty} \bar{\alpha}_t = 0$ by design of the diffusion process, it holds that $\lim_{t \to T} \frac{1-\bar{\alpha}_t}{\bar{\alpha}_t} = \lim_{t \to \infty} \frac{1-\bar{\alpha}_t}{\bar{\alpha}_t} = \infty$. This formally translates to the following condition:

$$\forall \epsilon > 0 \; \exists \tau\left(\epsilon\right) \in \mathbb{N} : \forall t > \tau \Rightarrow \frac{1-\bar{\alpha}_t}{\bar{\alpha}_t} \geq \epsilon. \tag{28}$$

Selection of $\epsilon = \frac{1+w}{1-w}\left(4 + \min\left(M/N, 1\right)\right) \geq \frac{1+w}{1-w} \frac{\mathbb{E}_{\boldsymbol{y}|\boldsymbol{x}_0} \|\boldsymbol{x}_0 - \boldsymbol{A}^\mathsf{T} \boldsymbol{y}\|_2^2}{N} > 0$ into eq. (28) translates it to

$$\forall w \in (0,1) \; \exists \tau\left(w\right) \in \mathbb{N} : \forall t > \tau \Rightarrow \frac{1-\bar{\alpha}_t}{\bar{\alpha}_t} \geq \frac{\left(1-w^2\right) \mathbb{E}_{\boldsymbol{y}|\boldsymbol{x}_0} \|\boldsymbol{x}_0 - \boldsymbol{A}^\mathsf{T} \boldsymbol{y}\|_2^2}{\left(1-w\right)^2 N}, \tag{29}$$

which concludes the proof given the equivalence of eq. (27) and eq. (16).

D Proof of Lemma 3.2

We prove the correctness of the transition kernel formula by induction on k.

Base case. First, we focus on the case $k = 1$, for which eq. (8) should match the transition probability from eq. (3). Indeed, from eq. (10), it holds:

$$\sum_{i=0}^{k-1} \Gamma_{t-i-1}^{t-k+1} \gamma_{t-i} = \Gamma_{t-1}^{t} \gamma_t = \gamma_t = \frac{\sqrt{\bar{\alpha}_{t-1}} \beta_t}{1-\bar{\alpha}_t}, \tag{30}$$

$$\Gamma_t^{t-k+1} = \Gamma_t^t = \frac{\sqrt{\alpha_t}\left(1-\bar{\alpha}_{t-1}\right)}{1-\bar{\alpha}_t}, \tag{31}$$

$$\sum_{i=0}^{k-1} \left(\Gamma_{t-i-1}^{t-k+1}\right)^2 \sigma_{t-i}^2 = \left(\Gamma_{t-1}^{t}\right)^2 \sigma_t^2 = \sigma_t^2. \tag{32}$$

Substituting these results in eq. (9) and then in eq. (8) leads us to eq. (3).

Induction Step. Let the induction hypothesis from eq. (8) to be valid for some k, such that $t - k \in (0, T]$. Then, we need to show that eq. (8) is also valid for $k+1$.

We note, that from eq. (3) we have $\boldsymbol{x}_{t-k-1} = \boldsymbol{\mu}_{t-k}\left(\boldsymbol{x}_{t-k}, \boldsymbol{x}_0\right) + \sigma_{t-k} \boldsymbol{\epsilon}_1$, where $\boldsymbol{\epsilon}_1 \sim \mathcal{N}\left(\boldsymbol{0}, \boldsymbol{I}_N\right)$. At the same time, since eq. (8) is valid for k by the induction hypothesis, we have that $\boldsymbol{x}_{t-k} =$

$\boldsymbol{\mu}_{t,k}\left(\boldsymbol{x}_t, \boldsymbol{x}_0\right) + \sigma_{t,k}\boldsymbol{\epsilon}_2$, where $\boldsymbol{\epsilon}_2 \sim \mathcal{N}\left(\mathbf{0}, \boldsymbol{I}_N\right)$. Combining these two equations with eqs. (3), (8) and (9), we get:

$$\boldsymbol{x}_{t-k-1} = \frac{\sqrt{\bar{\alpha}_{t-k-1}}\beta_{t-k}}{1-\bar{\alpha}_{t-k}}\boldsymbol{x}_0 + \frac{\sqrt{\alpha_{t-k}}\left(1-\bar{\alpha}_{t-k-1}\right)}{1-\bar{\alpha}_{t-k}}\boldsymbol{x}_{t-k} + \sigma_{t-k}\boldsymbol{\epsilon}_1 \tag{33}$$

$$= \boldsymbol{x}_0\left(\frac{\sqrt{\alpha_{t-k}}\left(1-\bar{\alpha}_{t-k-1}\right)}{1-\bar{\alpha}_{t-k}}\sum_{i=0}^{k-1}\Gamma_{t-i-1}^{t-k+1}\gamma_{t-i} + \frac{\sqrt{\bar{\alpha}_{t-k-1}}\beta_{t-k}}{1-\bar{\alpha}_{t-k}}\right) + \boldsymbol{x}_t\frac{\sqrt{\alpha_{t-k}}\left(1-\bar{\alpha}_{t-k-1}\right)}{1-\bar{\alpha}_{t-k}}\Gamma_t^{t-k+1}$$

$$+\sqrt{\frac{\alpha_{t-k}\left(1-\bar{\alpha}_{t-k-1}\right)^2}{\left(1-\bar{\alpha}_{t-k}\right)^2}\sum_{i=0}^{k-1}\left(\Gamma_{t-i-1}^{t-k+1}\right)^2\sigma_{t-i}^2}\boldsymbol{\epsilon}_2 + \sigma_{t-k}\boldsymbol{\epsilon}_1.$$

To move further, we first prove that $\forall i, j \in \mathbb{N}^+, i \geq j$, Γ_i^j from eq. (10) can be decomposed as: $\Gamma_i^j = \Gamma_j^j\Gamma_i^{j+1}$. This is easy to show by direct substitution, that is:

$$\Gamma_j^j\Gamma_i^{j+1} = \sqrt{\alpha_j}\frac{1-\bar{\alpha}_{j-1}}{1-\bar{\alpha}_j}\sqrt{\prod_{n=j+1}^{i}\alpha_n}\frac{1-\bar{\alpha}_j}{1-\bar{\alpha}_i}\sqrt{\alpha_j\prod_{n=j+1}^{i}\alpha_n}\frac{1-\bar{\alpha}_{j-1}}{1-\bar{\alpha}_i} = \sqrt{\prod_{n=j}^{i}\alpha_n}\frac{1-\bar{\alpha}_{j-1}}{1-\bar{\alpha}_i} = \Gamma_i^j. \tag{34}$$

We additionally note that it holds:

$$\frac{\sqrt{\alpha_{t-k}}\left(1-\bar{\alpha}_{t-k-1}\right)}{1-\bar{\alpha}_{t-k}} = \Gamma_{t-k}^{t-k}, \tag{35}$$

$$\frac{\sqrt{\bar{\alpha}_{t-k-1}}\beta_{t-k}}{1-\bar{\alpha}_{t-k}} = \gamma_{t-k} = \Gamma_{t-k-1}^{t-k}\gamma_{t-k} \tag{36}$$

$$\sigma_{t-k}^2 = \left(\Gamma_{t-k-1}^{t-k}\right)^2\sigma_{t-k}^2, \tag{37}$$

as $\Gamma_{t-k-1}^{t-k} = 1$, since $t-k-1 < t-k$.

First, we compute the multiplier of $\boldsymbol{x}_t$ in eq. (33). Since $t \geq t-k+1 \ \ \forall k \in \mathbb{N}^+$, we can use eq. (34) to derive the following result:

$$\frac{\sqrt{\alpha_{t-k}}\left(1-\bar{\alpha}_{t-k-1}\right)}{1-\bar{\alpha}_{t-k}}\Gamma_t^{t-k+1} = \Gamma_{t-k}^{t-k}\Gamma_t^{t-k+1} = \Gamma_t^{t-k}, \tag{38}$$

where we have additionally utilized the result of eq. (35). Next, we simplify the multiplier of $\boldsymbol{x}_0$ from eq. (33). We divide the sum inside this multiplier into two parts:

$$\sum_{i=0}^{k-1}\Gamma_{t-i-1}^{t-k+1}\gamma_{t-i} = \sum_{i=0}^{k-2}\Gamma_{t-i-1}^{t-k+1}\gamma_{t-i} + \Gamma_{t-k}^{t-k+1}\gamma_{t-k+1} = \sum_{i=0}^{k-2}\Gamma_{t-i-1}^{t-k+1}\gamma_{t-i} + \gamma_{t-k+1}, \tag{39}$$

where to separate the γ_{t-k+1} term we have used the lower case of Γ_i^j from eq. (10). Here we additionally note, that if $k=1$, then the remaining sum $\sum_{i=0}^{k-2}\Gamma_{t-i-1}^{t-k+1}$ becomes zero. If $k>1$, then for all the terms of this sum it holds $t-i-1 \geq t-k+1$. Combining this with the results of eqs. (34) to (36), we can compute the multiplier of $\boldsymbol{x}_0$ from eq. (33) as:

$$\frac{\sqrt{\alpha_{t-k}}\left(1-\bar{\alpha}_{t-k-1}\right)}{1-\bar{\alpha}_{t-k}}\sum_{i=0}^{k-1}\Gamma_{t-i-1}^{t-k+1}\gamma_{t-i} + \frac{\sqrt{\bar{\alpha}_{t-k-1}}\beta_{t-k}}{1-\bar{\alpha}_{t-k}} = \sum_{i=0}^{k-2}\Gamma_{t-k}^{t-k}\Gamma_{t-i-1}^{t-k+1}\gamma_{t-i} + \Gamma_{t-k}^{t-k}\gamma_{t-k+1}$$

$$+\Gamma_{t-k-1}^{t-k}\gamma_{t-k} = \sum_{i=0}^{k-2}\Gamma_{t-i-1}^{t-k}\gamma_{t-i} + \left[\Gamma_{t-i-1}^{t-k}\gamma_{t-i}\right]_{i=k-1} + \left[\Gamma_{t-i-1}^{t-k}\gamma_{t-i}\right]_{i=k} = \sum_{i=0}^{k}\Gamma_{t-i-1}^{t-k}\gamma_{t-i}. \tag{40}$$

To simplify the remaining part of eq. (33), which involves the terms with the Gaussian noise vectors $\boldsymbol{\epsilon}_1$ and $\boldsymbol{\epsilon}_2$, we utilize the fact that $\forall a, b \geq 0, \ \ \boldsymbol{\epsilon}_1, \boldsymbol{\epsilon}_2 \sim \mathcal{N}\left(\mathbf{0}, \boldsymbol{I}_N\right)$ it holds: $\sqrt{a}\boldsymbol{\epsilon}_1 + \sqrt{b}\boldsymbol{\epsilon}_2 \sim$

$\mathcal{N}(\mathbf{0}, (a+b)\boldsymbol{I}_N)$. Based on this, we can write:

$$\sqrt{\frac{\alpha_{t-k}\left(1-\bar{\alpha}_{t-k-1}\right)^2}{\left(1-\bar{\alpha}_{t-k}\right)^2}\sum_{i=0}^{k-1}\left(\Gamma_{t-i-1}^{t-k+1}\right)^2\sigma_{t-i}^2}\boldsymbol{\epsilon}_2+\sigma_{t-k}\boldsymbol{\epsilon}_1\sim\mathcal{N}\left(\mathbf{0},\hat{\sigma}_{t,k+1}^2\boldsymbol{I}\right), \tag{41}$$

where, if we additionally use the results of eqs. (35) and (37), we have that:

$$\hat{\sigma}_{t,k+1}^2=\left(\Gamma_{t-k}^{t-k}\right)^2\sum_{i=0}^{k-1}\left(\Gamma_{t-i-1}^{t-k+1}\right)^2\sigma_{t-i}^2+\left(\Gamma_{t-k-1}^{t-k}\right)^2\sigma_{t-k}^2. \tag{42}$$

Now, we use the same strategy as in eq. (39) and eq. (40) to simplify this expression:

$$\hat{\sigma}_{t,k+1}^2=\sum_{i=0}^{k-2}\left(\Gamma_{t-k}^{t-k}\Gamma_{t-i-1}^{t-k+1}\right)^2\sigma_{t-i}^2+\left(\Gamma_{t-k}^{t-k}\right)^2\sigma_{t-k+1}^2+\left(\Gamma_{t-k-1}^{t-k}\right)^2\sigma_{t-k}^2=\sum_{i=0}^{k}\left(\Gamma_{t-i-1}^{t-k}\right)^2\sigma_{t-i}^2. \tag{43}$$

From eqs. (38), (40) and (43), we get:

$$\boldsymbol{x}_{t-k-1}\sim\mathcal{N}\left(\hat{\boldsymbol{\mu}}_{t,k+1}\left(\boldsymbol{x}_t,\boldsymbol{x}_0\right),\hat{\sigma}_{t,k+1}^2\boldsymbol{I}_N\right), \tag{44}$$

where

$$\hat{\boldsymbol{\mu}}_{t,k+1}\left(\boldsymbol{x}_t,\boldsymbol{x}_0\right)=\sum_{i=0}^{k}\Gamma_{t-i-1}^{t-k}\gamma_{t-i}\boldsymbol{x}_0+\Gamma_t^{t-k}\boldsymbol{x}_t, \tag{45}$$

$$\hat{\sigma}_{t,k+1}^2=\sum_{i=0}^{k}\left(\Gamma_{t-i-1}^{t-k}\right)^2\sigma_{t-i}^2. \tag{46}$$

By direct substitution of $k=k+1$ into eq. (9) it is easy to show that

$$\begin{aligned}\hat{\boldsymbol{\mu}}_{t,k+1}\left(\boldsymbol{x}_t,\boldsymbol{x}_0\right)&=\boldsymbol{\mu}_{t,k+1}\left(\boldsymbol{x}_t,\boldsymbol{x}_0\right),\\ \hat{\sigma}_{t,k+1}^2&=\sigma_{t,k+1}^2,\end{aligned}$$

which implies that the induction hypothesis from eq. (8) is valid for $k+1$. This completes the proof of the induction step and combined with with the base case it proves by induction the validity of eq. (8) for every feasible k.

E Comparative Analysis: Proposed Accelerated Sampling and Prior work

E.1 Conceptual Difference

Using our notation, both [13] and [52] propose to start the reverse process from a timestep τ and a noisy version $\boldsymbol{x}_\tau$ of the initial estimate of $\boldsymbol{x}_0$, which we denote by $\mathbb{E}\left[\boldsymbol{x}_0|\boldsymbol{y}\right]$. The main conceptual difference of our approach is that in these cases $\boldsymbol{x}_\tau$ is obtained using the forward diffusion process, while in our case we end up in $\boldsymbol{x}_\tau$ using the reverse process. The initial motivation for our proposed approach is also different. In particular, while we motivate our procedure from a probabilistic viewpoint and propose to approximate the conditional score function as a composition of three functions, the authors in [13] base their strategy on the contrastive property of reverse SDEs, while the authors in [52] use the re-projection of unrealistic images to the manifold of natural images in the noisy latent space.

E.2 Technical Difference

Given that in our work we consider the standard DDPM realization of diffusion process (VP-SDE), we will explain the existing differences under this scenario. The authors of [13] and [52] propose to parameterize $\boldsymbol{x}_\tau$ as

$$\boldsymbol{x}_\tau=\sqrt{\bar{\alpha}_\tau}\mathbb{E}\left[\boldsymbol{x}_0|\boldsymbol{y}\right]+\sqrt{1-\bar{\alpha}_\tau}\boldsymbol{z},\ \ \boldsymbol{z}\sim\mathcal{N}\left(\mathbf{0},\boldsymbol{I}\right).$$

In contrast, in our case by using Eq.(10) we adopt the following parametrization:

$$\boldsymbol{x}_\tau = \sum_{i=0}^{T-\tau-1} \Gamma_{T-i-1}^{\tau+1} \gamma_{T-i} \mathbb{E}\left[\boldsymbol{x}_0 | \boldsymbol{y}\right] + \Gamma_T^{\tau+1} \boldsymbol{x}_T + \sqrt{\sum_{i=0}^{T-\tau-1} \left(\Gamma_{T-i-1}^{\tau+1}\right)^2 \sigma_{T-i}^2} \boldsymbol{z},$$

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

Thus, our parametrization is more general and it is possible to show by induction, that under certain conditions it leads to the exact same $\boldsymbol{x}_\tau$ as in [13] and [52].

Finally, to experimentally demonstrate that our approach exhibits certain benefits compared to the ones described in [13] and [52], we conducted additional comparisons for the SISR problem between the different sampling strategies (see Table 8). From these results it is clear that our proposed strategy works better in practice and leads to superior results both in terms of fidelity and perceptual quality.

Table 8: Comparison of the proposed acceleration scheme and prior works [13, 52] for the SISR task.

Acceleration Strategy	PSNR ↑	SSIM ↑	LPIPS ↓	TOPIQ$_\Delta$ ↓	NFE ↓
Ours	**28.12**	**0.793**	**0.140**	**0.002**	51
[13] and [52]	28.05	0.783	0.142	0.016	51

F Computational Cost Analysis

In this section, we calculate the computational cost for each diffusion-based method in terms of TFLOPs. The input image size for all competing approaches is 720p (1280 x 720).

Table 9: Computational cost of the proposed and existing diffusion-based methods for the Dynamic Scene Deblurring task with 720p input resolution.

Method	TFLOP (equation)	TFLOP Total ↓
DvSR [78]	1.2×NFE + 4.8	604.8
icDPM [59]	4.8×NFE + 5.2	2405.2
InDI [15]	4.8×NFE	48.0
Ours	4.3×NFE + 1.9	**23.4**

The proper way to interpret equations in Table 9 is as follows: $\text{TFLOP}_{\text{total}} = x \times N + y$, where x is the TFLOP complexity for a single backbone pass within the diffusion process, N is the total number of neural function evaluations (NFEs) per sampling process, and is the complexity of sub-modules that have to be run once per image (e.g. Image Restoration network in our method, pre-processing net for icDPM [59] and DvSR [78]). Based on these results, we observe that the computational cost of our method is significantly lower compared to our diffusion-based competitors.

G Proposed one-step acceleration with/without DDIM

Utilizing our one-step acceleration, we bypass steps from $t = T$ to $t = \tau$, allowing us to either straightforwardly execute the remaining $t = \tau$ steps or apply any existing acceleration strategies. To demonstrate that our one-step acceleration is complementary to existing accelerated sampling strategies, we combined the DDIM [66] acceleration technique with our one-step acceleration for the single-image super-resolution (SISR) task (refer to Table 10). We notice that no significant quantitative/qualitative difference after applying this acceleration technique has been observed.

Table 10: Results for the proposed one-step acceleration with/without DDIM acceleration tested on SISR task

Method	PSNR ↑	SSIM ↑	LPIPS ↓	TOPIQ$_\Delta$ ↓	NFE ↓
Ours (with DDIM)	28.12	0.793	0.140	**0.002**	51
Ours w/o DDIM	**28.16**	**0.794**	**0.139**	0.007	251

H Training Procedure

Denoising Module. We use two denoising architectures, namely MIRNet-S and UDP [7] for our main experiments and ablation studies. MIRNet-S (15.95M parameters) is a lighter version of the original MIRNet (31.78M parameters) [81], where MSRB is decreased from 2 to 1. In contrast, UDP (11.78M parameters) [7] architecture is not modified and used as it is. Both models are trained for the Gaussian denoising task in the sRGB domain with input noise level $\tilde{\sigma} \in [0, 244.3]$. More specifically, for each element of batch the noise standard deviation is selected randomly using uniform sampling within this range. The batch size, training crop size, and initial learning rate are set to 8 (in total), 192×192, and $2 \cdot 10^{-4}$, respectively. Overall, all denoisers are trained for 1M iterations on 8 Ascend 910 AI accelerators using the Adam [32] optimizer with default parameters and a decaying learning rate scheduler: $lr_s = lr_0 * \gamma^{\lfloor s/2000 \rfloor}$, where $\gamma = 0.999$. We employ an MSE loss to train the denoising model and concatenate the noise level with the noisy image as an input to the denoiser same as [7].

IR Module. As mentioned in the main manuscript, we utilize existing models with publicly available pretrained parameters. For the tasks of burst JDD-SR, dynamic scene deblurring, and SISR we have used BSRT-Small [50], FFTFormer [34], and SwinIR [42], respectively. A description of each one of these IR networks, including their number of trainable parameters, is provided in Table 11.

Fusion Module. For each IR task under study, we follow the same protocol. We train the fusion module for $300K$ iterations with batch size of 128 (in total), and crop size of 256×256. The training takes place on 8 Ascend 910 AI accelerators, while the optimizer of choice is Adam [32] with default parameters and a learning rate scheduler $lr_s = lr_0 * \gamma^{\lfloor s/1000 \rfloor}$, where $\gamma = 0.99$. For each one of the studied IR tasks we train our Fusion module on a dedicated dataset. Specifically, we use the ZurichRaw2RGB [29] dataset for burst JDD-SR, GoPro [53] for dynamic scene deblurring, and DIV2K [1] for SISR. The selection of these specific datasets is motivated by the fact that they are widely used by all the competing methods for network training related to the IR tasks of interest. The detailed description of training data we used for each problem is provided below.

- **JDD-SR.** Following the same protocol as in [4, 37, 49, 5, 17, 50], we generate 46K burst sets from the training set of ZurichRAW2RGB. Each burst set, which contains 14 low-resolution images in the raw domain, is inferenced by BSRT-Small [50] and post-processed to produce an image in the sRGB domain. Then, these predictions are used as input to train **only** our Fusion module, while Denoising and IR networks are frozen.
- **Dynamic Scene Deblurring.** We follow the standard protocol for this problem and use the GoPro dataset for training. Specifically, we use 3214 pairs of clean and blurry 1280×720 images, out of which we have excluded the 1111 pairs reserved for evaluation purposes. In order to provide a fair comparison, we follow exactly the same setup as in [35, 85, 83, 34] and train **only** the fusion module using the provided GoPro training data.
- **SISR** We employ the well-known DIV2K [1] dataset for the SISR task. This dataset contains a set of 800 images of 2K resolution, which we used for training our Fusion module. Following the standard protocol, we use the additionally provided 100 2K images for evaluation.

I Fusion Module

The main goal of our proposed Fusion module is to predict the conditional expectation $\mathbb{E}[\boldsymbol{x}_0|\boldsymbol{y}, \boldsymbol{x}_t]$ given the estimates of $\mathbb{E}[\boldsymbol{x}_0|\boldsymbol{x}_t]$ and $\mathbb{E}[\boldsymbol{x}_0|\boldsymbol{y}]$, which are produced by the denoising and IR modules, respectively. To do so, our fusion network accepts as inputs the image estimates $\boldsymbol{x}_0^D, \boldsymbol{x}_0^{IR}$ and a timestep t. Its exact architecture is depicted in Figure 5 and it consists of two branches. The upper branch operates on $\boldsymbol{x}_0^D$, $\boldsymbol{x}_0^{IR}$ and produces the corresponding features $\boldsymbol{f}_1$, $\boldsymbol{f}_2$ using a single dense block [27] without a normalization layer. The lower branch encodes the timestep t into a vector of weights $\boldsymbol{w} \in (0, 1)^{n_f}$ using the sinusoidal positional encoding [71], followed by a two layer MLP, and a sigmoid function as the final activation. Then, we perform a weighted summation of $\boldsymbol{f}_1$, $\boldsymbol{f}_2$ in the feature space. Finally, two consequent dense blocks [27], with n_f channels each, followed by

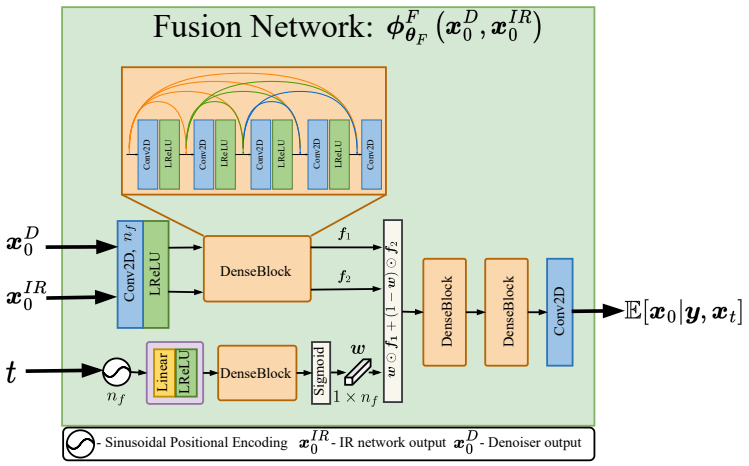

Figure 5: Detailed visualization of the proposed Fusion module.

Table 11: Number of parameters and source code of methods used as an IR module in our framework.

IR Modules	Parameters ↓	Code & Weights
BSRT-Small	4.92M	link
FFTFormer	16.56M	link
SwinIR	11.85M	link

a convolution layer produce the final estimate of $\mathbb{E}[x_0|y, x_t]$. Our proposed architecture has only 0.73M learnable parameters, which is significantly lower (7-22 times) compared to the Denoising and IR modules. As a result, the Fusion module requires less training time/resources and can be trained only on a small amount of problem-specific training data.

J Fusion Module Robustness

In order to assess the robustness/generalization ability of our method, we have conducted additional experiments where we evaluate the reconstruction quality achieved when the Fusion network, which was originally trained on the MIRNet-S and SwinIR pair, is combined with the following pairs of denosing and IR networks:

- The Fusion module is combined with the same pair of denoising and IR networks as those used during its training.
- The Fusion module is combined with a different IR network and the same denoising network as the one used during its training.
- The Fusion module is combined with a different denoising network and the same IR network as the one used during its training.
- The Fusion module is combined with different denoising and IR networks than the ones used during its training.

Table 12: The performance of Fusion module for different train/test pair scenarios for 4x SR task.

	Trained Pair	Tested Pair	PSNR↑	SSIM↑	LPIPS↓	TOPIQ$_\Delta$↓
-	Target		∞	1	0	0
1	MIRNet-S + SwinIR	MIRNet-S + SwinIR	28.12	0.793	0.140	0.002
2	MIRNet-S + SwinIR	MIRNet-S + RRDB	28.20	0.795	0.144	0.023
3	MIRNet-S + SwinIR	UDP + SwinIR	28.47	0.808	0.177	0.017
4	MIRNet-S + SwinIR	UDP + RRDB	28.46	0.807	0.183	0.025

From these experiments as shown in the table above, we observe that changing the denoising network to a less powerful one leads to a noticeable drop in terms of perceptual quality (30% in LPIPS) and

slightly blurrier results, which corresponds to a PSNR increase by 0.2dB. Conversely, altering the IR network has a minimal impact on both the reconstruction and perceptual metrics. In summary, under the particular IR task, the Fusion network evaluated with a different pair of denoising and IR modules demonstrates a good generalization ability and a robust behavior.

K Perception-Fidelity Trade-off

We have evaluated our method on the GoPro dataset (motion deblurring) for different τ, when the denoising and Fusion modules are activated. From Table 13, we observe that the perception-fidelity trade-off follows a similar trend to the one for the SISR 4x task. The main difference is that in this particular case the necessary reverse diffusion steps are less than in SISR.

Table 13: Perception-Distortion Trade-off on GoPro validation for dynamic scene deblurring task.

τ	**0**	**1**	**5**	**10**	**20**	**25**
PSNR, dB	34.21	34.02	33.72	33.52	33.23	33.14
LPIPS	0.071	0.057	0.053	0.052	0.052	0.052

L Information about Competing methods.

Below, we provide links to the code implementation and trained weights for all baselines used for comparison.

Table 14: Code and model weights of the competing methods for burst JDD-SR task.

Method	Link to Code	Link to Weights
DBSR	link	link
DeepRep	link	link
EBSR	link	link
BIPNet	link	link
BSRT	link	link

Table 15: Code and model weights of the competing methods for dynamic scene deblurring task.

Method	Link to Code	Link to Weights
HINet	link	link
MPRNet	link	link
MIMO-UNet+	link	link
NAFNet	link	link
Restormer	link	link
FFTFormer	link	link
DeblurGANv2	link	link

M Additional Results

In the section, we provide additional visual comparison of the proposed method with existing SOTA approaches for all IR tasks under study.

Table 16: Code and model weights of the competing methods for SISR task.

Method	Link to Code	Link to Weights
SRResNet	link	link
RRDB	link	link
SwinIR	link	link
LIIF	link	link
HAT	link	link
ESRGAN	link	link
HCFlow	link	link
SwinIR-GAN	link	link
LDM	link	link
SRDiff	link	link
IDM	link	link

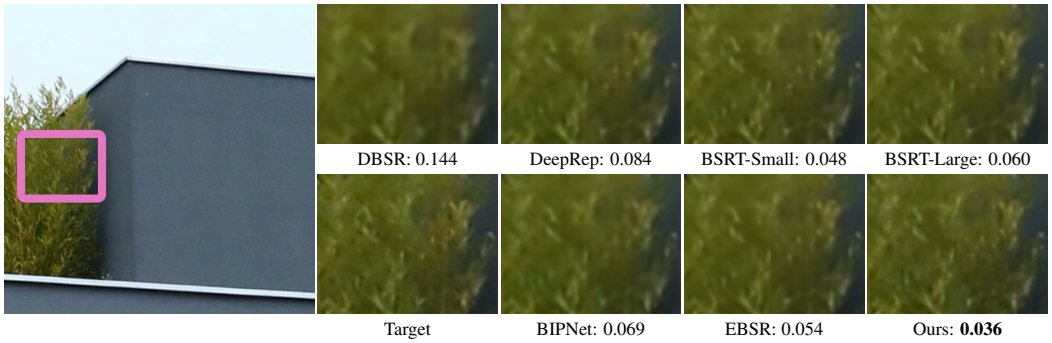

Figure 6: Visual comparison of our approach against competing methods on the Burst JDD-SR task (best viewed by zooming in). Every output image is accompanied by its LPIPS value.

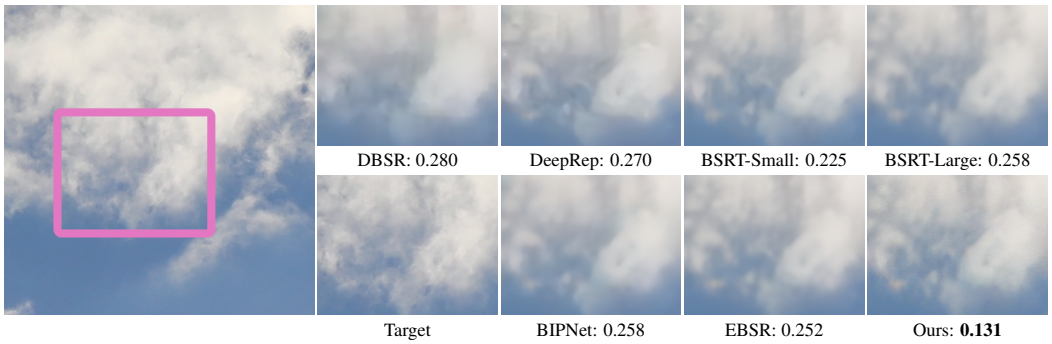

Figure 7: Visual comparison of our approach against competing methods on the Burst JDD-SR task (best viewed by zooming in). Every output image is accompanied by its LPIPS value.

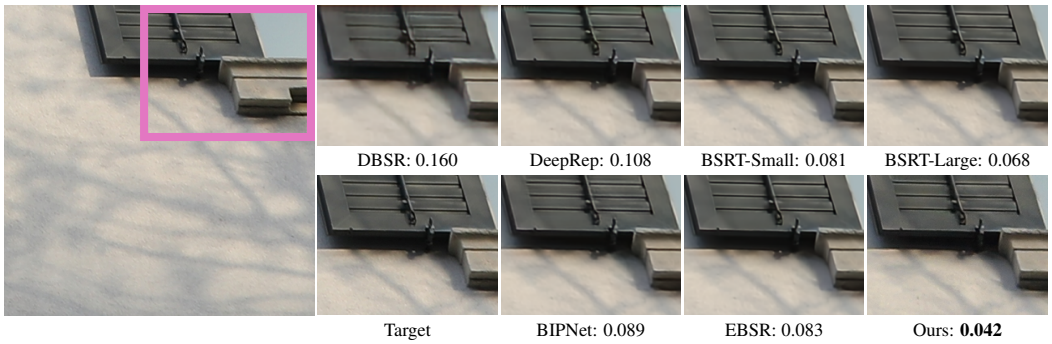

Figure 8: Visual comparison of our approach against competing methods on the Burst JDD-SR task (best viewed by zooming in). Every output image is accompanied by its LPIPS value.

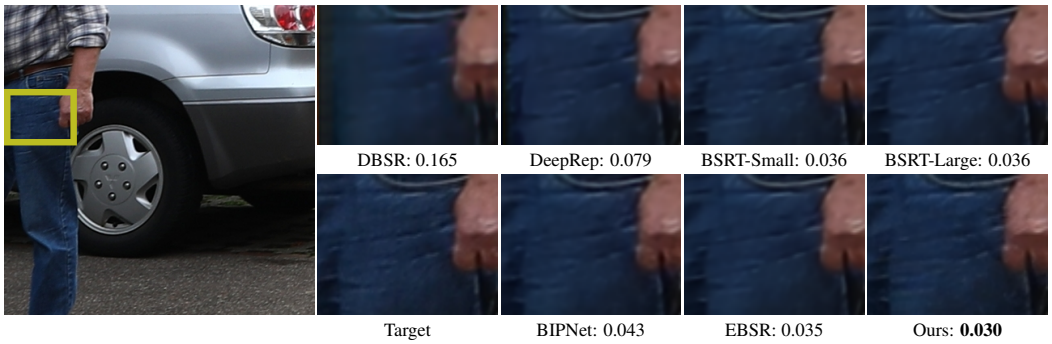

Figure 9: Visual comparison of our approach against competing methods on the Burst JDD-SR task (best viewed by zooming in). Every output image is accompanied by its LPIPS value.

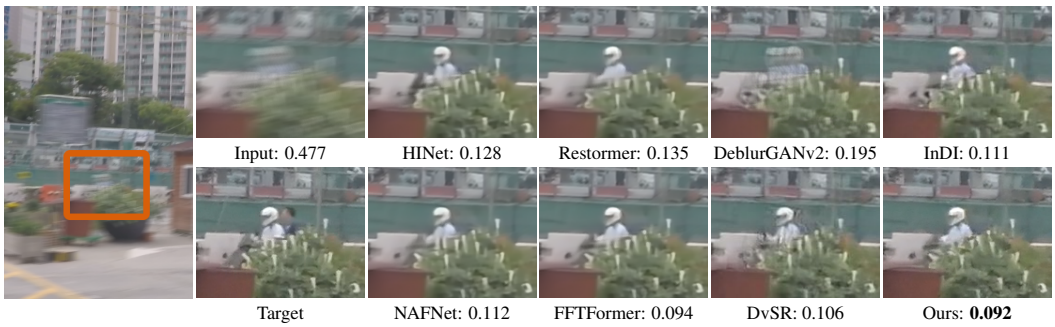

Figure 10: Visual comparison of our approach against competing methods on the GoPro test set for the task of dynamic scene deblurring (best viewed by zooming in). Every output image is accompanied by its LPIPS value.

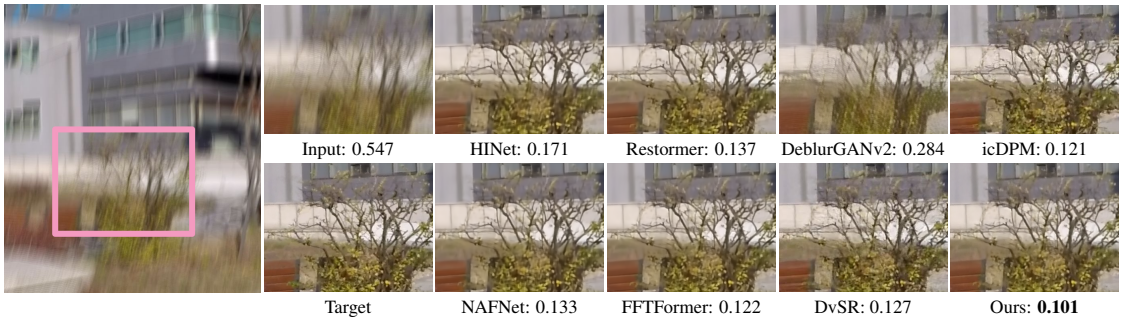

Figure 11: Visual comparison of our approach against competing methods on the GoPro test set for the task of dynamic scene deblurring (best viewed by zooming in). Every output image is accompanied by its LPIPS value.

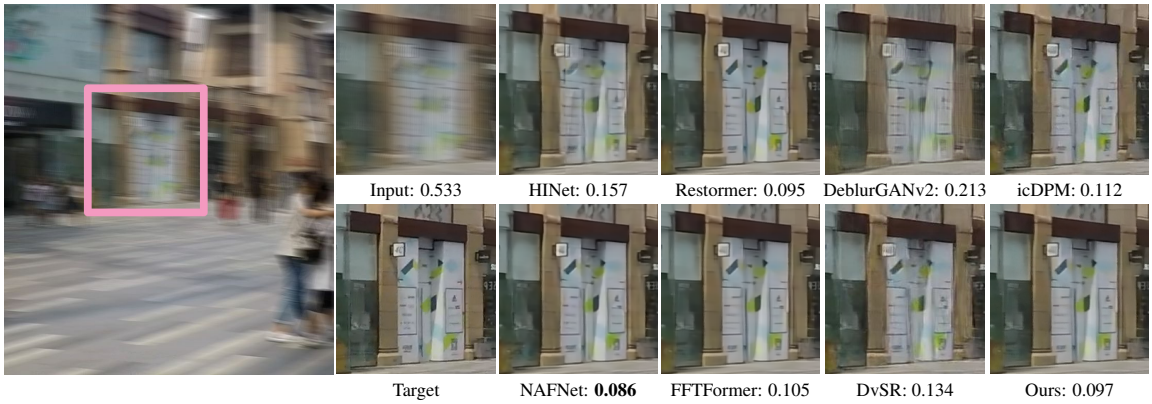

Figure 12: Visual comparison of our approach against competing methods on the HIDE test set for the task of dynamic scene deblurring (best viewed by zooming in). Every output image is accompanied by its LPIPS value.

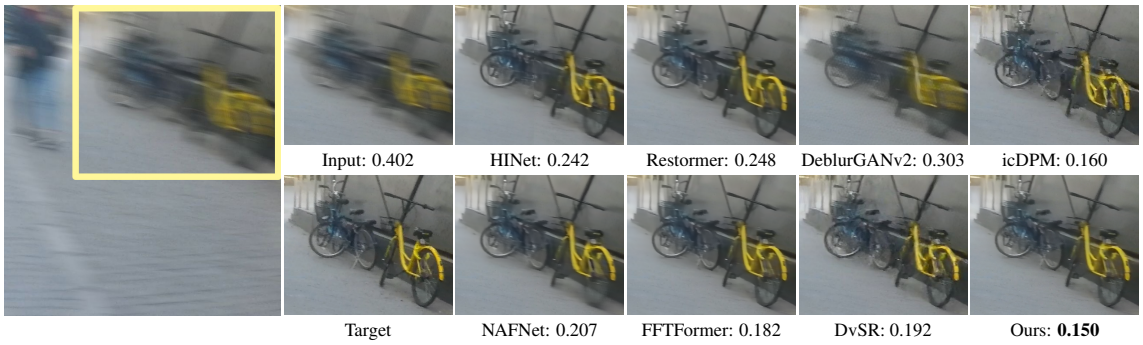

Figure 13: Visual comparison of our approach against competing methods on the HIDE test set for the task of dynamic scene deblurring (best viewed by zooming in). Every output image is accompanied by its LPIPS value.

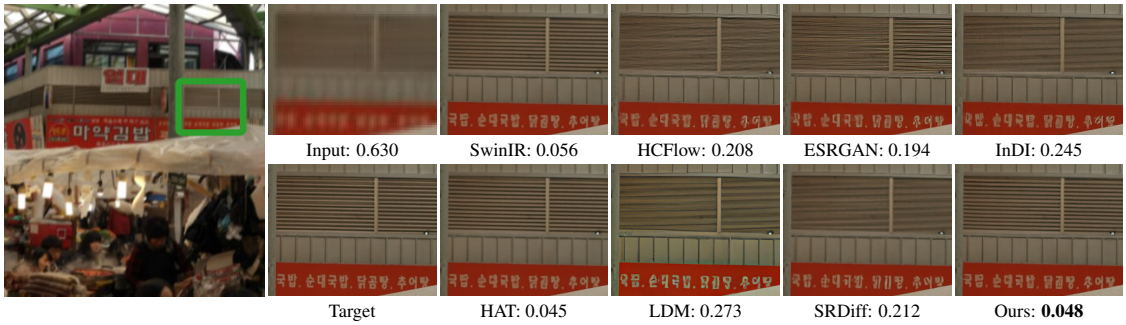

Figure 14: Visual comparison of our approach against competing methods on the DIV2K validation set for the task of $4\times$ SISR (best viewed by zooming in). Every output image is accompanied by its LPIPS value.

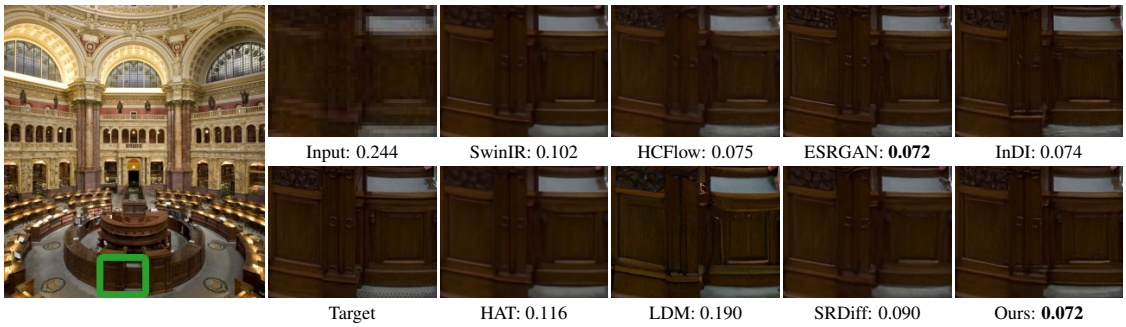

Figure 15: Visual comparison of our approach against competing methods on the DIV2K validation set for the task of $4\times$ SISR (best viewed by zooming in). Every output image is accompanied by its LPIPS value.

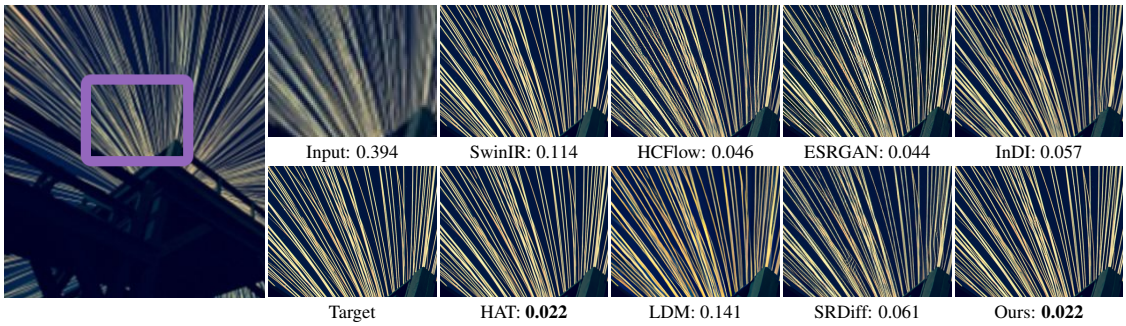

Figure 16: Visual comparison of our approach against competing methods on the DIV2K validation set for the task of $4\times$ SISR (best viewed by zooming in). Every output image is accompanied by its LPIPS value.

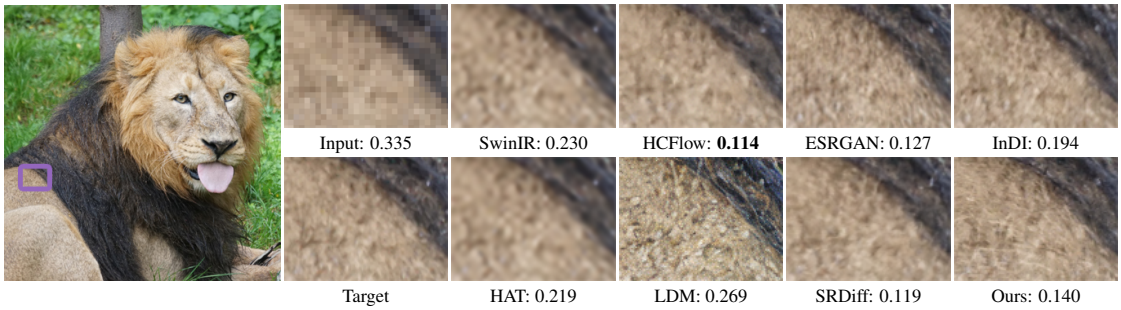

Figure 17: Visual comparison of our approach against competing methods on the DIV2K validation set for the task of $4\times$ super-resolution (best viewed by zooming in). Every output image is accompanied by its LPIPS value.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The claims made in the abstract and section 1 match the theoretical and experimental results provided in the next sections. We do not mention any aspirational claims in our work.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The limitations of our work are properly discussed in a dedicated paragraph 5.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: For all the theoretical results mentioned in the paper, in section 3 and in appendix we provide detailed descriptions, a full set of assumptions and complete proofs.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: In section 4 and appendices H and I we provide the complete and detailed description of our training and inference procedures, that enables the reproducibility of our results.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: All the data we used in our work is publicly available. Upon the acceptance of the paper we plan to release the inference code together with the trained models checkpoints.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: In section 4 and appendices H and I we provide the detailed description of our training and inference procedures.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: Following the common practices in the field of image restoration, we do not report the error bars, as the correct estimation of them is computationally expensive.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: For all the experiments the information on the computer resources is provided in appendix H.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Our research strictly follows the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: Besides the fact that in our work we train the conditional generative models, their limited capacity does not allow to generate anything that could be considered as a negative social impact. The proposed techniques are used to increase the perceptual appearance of the low-quality images and hardly pose any positive/negative societal impact.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA] .

    Justification: All the data used for training of our networks or pre-trained networks that we have utilized in our work is publicly available and widely known to be safe, so the paper poses no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: all the datasets used in the paper are publicly available, and in section 4 we properly cite the papers where such datasets were introduced.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.

- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA] .

    Justification: The submission does not release new assets. Upon acceptance we plan to release well documented inference code under the non-commercial usage licence.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA] .

    Justification: the paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA] .

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.