Deblurring via Stochastic Refinement

Jay Whang[†*] Mauricio Delbracio[‡] Hossein Talebi[‡] Chitwan Saharia[‡]
Alexandros G. Dimakis[†] Peyman Milanfar[‡]
[†]University of Texas at Austin [‡]Google Research

Abstract

Image deblurring is an ill-posed problem with multiple plausible solutions for a given input image. However, most existing methods produce a deterministic estimate of the clean image and are trained to minimize pixel-level distortion. These metrics are known to be poorly correlated with human perception, and often lead to unrealistic reconstructions. We present an alternative framework for blind deblurring based on conditional diffusion models. Unlike existing techniques, we train a stochastic sampler that refines the output of a deterministic predictor and is capable of producing a diverse set of plausible reconstructions for a given input. This leads to a significant improvement in perceptual quality over existing state-of-the-art methods across multiple standard benchmarks. Our predict-and-refine approach also enables much more efficient sampling compared to typical diffusion models. Combined with a carefully tuned network architecture and inference procedure, our method is competitive in terms of distortion metrics such as PSNR. These results show clear benefits of our diffusion-based method for deblurring and challenge the widely used strategy of producing a single, deterministic reconstruction.

1. Introduction

Image deblurring is a long-standing problem in computer vision. Various conditions such as moving objects, camera shakes, or an out-of-focus lens may contribute to blurring artifacts. Single image deblurring is a highly ill-posed inverse problem where multiple plausible sharp images could lead to the very same blurry observation. Nonetheless, most existing methods produce a single deterministic estimate of the clean image.

Traditional methods formulate deblurring as a variational optimization problem and find a solution that satisfies closeness to certain image and/or blur kernel prior [9,18,28,38,58]. With the emergence of deep learning, convolutional neural networks (CNNs) have become the de-facto standard for deblurring models [14,35,40,59,63,65,66,74]. Typically, these CNNs are trained with simulated sharp-blurry image pairs through supervised learning. Minimizing L_1 or L_2 pixel loss is perhaps the most widely adopted approach for training such models. These losses provide a straightforward learning objective and optimize for the popular PSNR

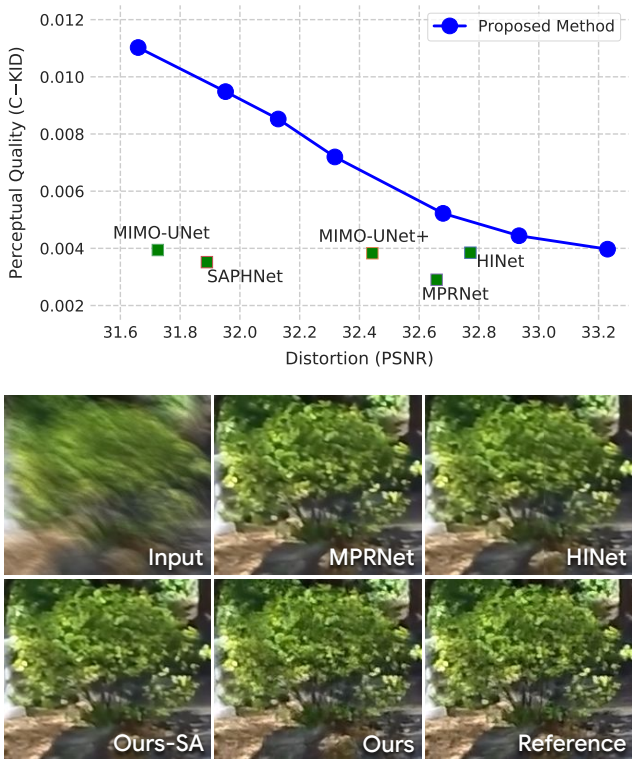

Figure 1. **Top:** Perception-Distortion (P-D) trade-off [5] of current state-of-the-art deblurring methods (top). Our method sets a new Pareto frontier in the P-D plot and allows us to traverse through the P-D curve using a single model without retraining or finetuning. **Bottom:** Samples from our method compared to other competitive methods. We include two extremes from our model – one optimized for perceptual quality ("Ours") and one for distortion using Sample Averaging ("Ours-SA"). These correspond to the two end points of the P-D curve. For the ease of interpretation, we used negative Kernel Inception Distance [4] (C − KID for a constant C)

*This work was done during an internship at Google Research.

(peak signal-to-noise-ratio) metric. Unfortunately, PSNR and other distortion metrics are well-known to only partially correspond to human perception [5, 17, 19] and can actually lead to algorithms with visibly lower quality in the reconstructed images. To alleviate this problem, recent works introduced additional loss terms [17, 21, 34, 44, 45] that seek to improve the quality of generated images under metrics that represent human perception more reliably. Training networks to go from corrupted images to a known ground truth in a supervised way belongs in the family of end-to-end methods [50]. These methods perform very well in-distribution, but can be quite fragile to distributional shifts or changes in the corruption process [25, 50].

A second body of work has focused on using deep generative models to solve inverse problems [6]. For deblurring, Generative Adversarial Networks (GANs) [22] have been successfully applied with competitive performance [3, 34, 35]. GAN-based restoration methods train the deblurring network with an adversarial loss to make the restored images more perceptually plausible. However the proposed methods so far have been deterministic, and adversarial losses often introduce artifacts not present in the original clean image, leading to large distortion (*e.g.* [42] for super-resolution).

In this work, we adopt a different perspective and view deblurring as a conditional generative modeling task, where we seek to generate diverse samples from the posterior distribution. Specifically, we introduce a **"predict-and-refine" conditional diffusion model**, where a deterministic data-adaptive predictor is jointly trained with a stochastic sampler that refines the output of the said predictor (see Fig. 2).

Our predict-and-refine approach enables more efficient sampling compared to the standard diffusion model. This formulation also naturally leads to a stochastic model capable of producing realistic images without sacrificing pixel-level distortion. To the best of our knowledge, this is the first blind deblurring technique that leverages a deep generative model and is capable of producing *diverse* samples.

Overall, our method produces a variety of plausible and photo-realistic results, while achieving state-of-the-art performance under many quantitative metrics in terms of both distortion and perceptual quality across multiple standard datasets. In addition, by aggregating a different number of generated deblurred samples, our framework allows us to conveniently traverse the Perception-Distortion curve [5, 19] as shown in Fig. 1, without any expensive retraining or finetuning. These results show clear benefits of stochastic diffusion-based methods for deblurring and challenge the currently dominant strategy of producing deterministic reconstructions.

2. Related Work

The goal of image deblurring is to generate a plausible reconstruction of the unobserved sharp, clean image $\boldsymbol{x}$ from a blurry input $\boldsymbol{y}$. Deblurring techniques differ in what they aim to obtain. For example, one could try to directly sample from the posterior $p(\boldsymbol{x} \mid \boldsymbol{y})$. Another viable option is to compute a point-estimate such as the conditional mean $\mathbb{E}[\boldsymbol{x} \mid \boldsymbol{y}]$ or the *maximum a posteriori* estimate $\arg\max_{\boldsymbol{x}} p(\boldsymbol{x} \mid \boldsymbol{y})$.

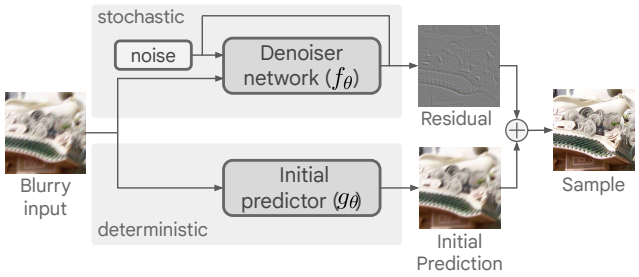

Figure 2. Diagram describing our dual-network architecture. The initial predictor produces the deterministic candidate for the denoiser network, which then models the residual.

Deblurring through point estimates. Traditional deblurring methods formulate the problem as one of blind deconvolution [9, 10, 16, 18, 28, 36, 38, 58, 72, 76]. In this setup, the blur is generally modeled as a noisy linear operator acting on the clean image. While the exact values of the blur operator are not assumed to be known, one can enforce some prior distribution on the blur and the sharp image and try to find the most likely solution.

Alternatively, many recent methods adopt an end-to-end approach where a deep neural network is trained to directly produce a point estimate [8, 11, 14, 20, 34, 35, 48, 52, 53, 62, 64, 65, 71]. These methods generally rely on pairs of blurry-sharp images as training data and cast the deblurring problem as a supervised regression task. Much of the efforts have gone into developing specialized network architectures and loss functions to achieve better pixel-level reconstruction metrics such as PSNR or SSIM [68]. For example, MIMO-UNet [14] proposed an architecture that facilitates information flow across different image resolutions in a multi-scale U-Net [55]. Another work HINet [11] introduced Half Instance Normalization [67], which can be used as a building block for image restoration networks. MPRNet [73] presented an improved multi-stage architecture designed to incorporate both high-level global features as well as local details.

Issue of *regression to the mean*. While the aforementioned approaches lead to state-of-the-art PSNR, they share the limitation that they can only produce a deterministic output. This is at odds with the nature of blind image deblurring, which is an inherently ill-posed inverse problem with *multiple* valid solutions for a single input. In fact, the current trend of developing point-estimators that directly minimize a distortion loss suffers from the problem of "regression to the mean". If there are multiple possible clean images that correspond to the blurry input, the optimal reconstruction

according to the given loss function will be an **average** of them. Consequently, the resultant deterministic reconstruction often lacks details as it learns to produce the average of all possible solutions at best.

Diverse image restoration. One way to circumvent the *regression to the mean* phenomenon is to avoid point estimations and directly learn to generate samples from the posterior distribution [30–32, 49]. While techniques based on adversarial training have been explored for blind deblurring [34, 35], in general they are not trained to produce multiple samples. Additionally, non-reference based adversarial losses can introduce significant hallucinations and distortions [15].

Likelihood-based deep generative models such as Variational Autoencoders [51], Normalizing Flows [42, 43], and Diffusion Probabilistic Models (DPMs) [39, 56] have also been successfully applied to other image enhancement tasks such as super-resolution, where a diverse set of candidates can be generated from the learned posterior [51]. Compared to point estimates, solving imaging inverse problems by sampling from the posterior has additional benefits such as uncertainty quantification [31, 32, 70], near-optimal sample complexity [27] and better fairness guarantees [26].

3. Diffusion Probabilistic Models

Diffusion probabilistic model [24, 60] is a latent variable model specified by a T-step Markov chain $(\boldsymbol{x}_0, \boldsymbol{x}_1, \dots, \boldsymbol{x}_T)$ called the *diffusion process*. It starts from a clean data sample $\boldsymbol{x}_0 \in \mathbb{R}^d$ and repeatedly injects Gaussian noise according to the transition kernel $q(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1})$ as follows:

$$q(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1}) \triangleq \mathcal{N}(\boldsymbol{x}_t; \sqrt{\alpha_t}\boldsymbol{x}_{t-1}, (1-\alpha_t)\boldsymbol{I}_d), \quad (1)$$

where $\alpha_t \in (0,1)$ for all $t = 1, \dots, T$. The noise schedule $\boldsymbol{\alpha}_{1:T} \triangleq (\alpha_1, \dots, \alpha_T)$ is a hyperparameter that controls the variance of noise added at each step. The latent variables $\boldsymbol{x}_{1:T}$ have the same dimensionality as the original data sample $\boldsymbol{x}_0$.

While this particular choice of diffusion process may seem arbitrary, it results in closed-form expressions for the following distributions: the marginal[1] distribution $q(\boldsymbol{x}_t \mid \boldsymbol{x}_0)$ and the *reverse diffusion step* $q(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t, \boldsymbol{x}_0)$. Writing $\bar{\alpha}_t \triangleq \prod_{j=1}^{t} \alpha_j$, we get

$$q(\boldsymbol{x}_t \mid \boldsymbol{x}_0) = \mathcal{N}(\boldsymbol{x}_t; \sqrt{\bar{\alpha}_t}\boldsymbol{x}_0, (1-\bar{\alpha}_t)\boldsymbol{I}_d) \quad (2)$$

$$q(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t, \boldsymbol{x}_0) = \mathcal{N}(\boldsymbol{x}_{t-1}; \boldsymbol{\mu}_t(\boldsymbol{x}_t, \boldsymbol{x}_0), \beta_t \boldsymbol{I}_d), \quad (3)$$

where $\boldsymbol{\mu}_t(\boldsymbol{x}_t, \boldsymbol{x}_0)$ and β_t are quantities that depend on $\boldsymbol{x}_t, \boldsymbol{x}_0$ and $\boldsymbol{\alpha}_{1:T}$. Their full expressions and derivations are included in the supplementary material.

The marginal distribution in Eq. (2) allows us to sample a partially noisy image $\boldsymbol{x}_t$ at an arbitrary time step, and the reverse diffusion step in Eq. (3) is a stochastic denoising procedure that tells us how to reverse a single diffusion step by sampling a slightly less noisy image $\boldsymbol{x}_{t-1}$ from $\boldsymbol{x}_t$. The ability to sample from arbitrary marginals is important to make training of a DPM practical, as the training objective relies on it (see Eq. (5)).

We note that the diffusion process defined here has no learnable parameter. It is a fixed process that gradually destroys the original signal $\boldsymbol{x}_0$ and produces $\boldsymbol{x}_T$ that looks indistinguishable from pure Gaussian noise given a sufficiently large T. Thus, if we could apply the reverse diffusion step T times starting from pure Gaussian noise, we would obtain a clean sample $\boldsymbol{x}_0$. However this is not possible because the reverse diffusion step itself requires access to $\boldsymbol{x}_0$, which is exactly what we are trying to generate.

Reverse process and denoiser network. A key component of DPM is the *denoiser network* f_θ that tries to estimate $\boldsymbol{x}_0$ from the partially noisy image $\boldsymbol{x}_t$. With it, we can apply the reverse diffusion step without knowing $\boldsymbol{x}_0$ by using the estimate $f_\theta(\boldsymbol{x}_t, t)$ in place of $\boldsymbol{x}_0$:

$$p_\theta(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t) \triangleq q(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t, f_\theta(\boldsymbol{x}_t, t)) \quad (4)$$

This defines a Markov chain that runs backwards in time from $\boldsymbol{x}_T$ to $\boldsymbol{x}_0$, which we call the *reverse process*. The goal of DPM is to train f_θ to make $p_\theta(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t)$ as close to the true reverse diffusion step $q(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t, \boldsymbol{x}_0)$ as possible. This is done by optimizing f_θ to maximize the variational lower bound of the marginal likelihood $\log p_\theta(\boldsymbol{x})$.

In practice, we use an alternative parametrization of f_θ proposed by [24] that instead predicts the Gaussian noise $\boldsymbol{\epsilon}$ that deterministically relates $\boldsymbol{x}_t$ and $\boldsymbol{x}_0$ via Equation (2). Specifically, we write $\boldsymbol{x}_t = \sqrt{\bar{\alpha}_t}\boldsymbol{x}_0 + (1-\bar{\alpha}_t)\boldsymbol{\epsilon}$ for $\boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_d)$ and train f_θ to predict $\boldsymbol{\epsilon}$.

Continuous noise level. Chen *et al*. [12] proposes a modified formulation based on a continuous noise level $\bar{\alpha}$, which we also adopt. An important property of this formulation is that it allows us to sample from the model using a noise schedule $\boldsymbol{\alpha}_{1:T}$ different from the one used during training. This flexibility enables us to control the trade-off between the distortion and the perceptual quality of generated samples without having to retrain the model, as we show later.

Conditional DPM. So far we have defined a DPM that is trained to model the *unconditional* data distribution. For conditional models that must estimate $p(\boldsymbol{x} \mid \boldsymbol{y})$, we make f_θ accept $\boldsymbol{y}$ as the conditioning input, as was done in [13, 56]. This way, the iterative denoising procedure becomes dependent on $\boldsymbol{y}$. The final training objective is:

$$L_{\text{Base}}(\theta) = \mathbb{E}\left\|\boldsymbol{\epsilon} - f_\theta(\sqrt{\bar{\alpha}}\boldsymbol{x}_0 + \sqrt{1-\bar{\alpha}}\boldsymbol{\epsilon}, \bar{\alpha}, \boldsymbol{y})\right\|_1, \quad (5)$$

where the expectation is over $\boldsymbol{y}, \boldsymbol{x}_0, \bar{\alpha}$, and $\boldsymbol{\epsilon}$.

Sampling from a DPM. As mentioned earlier, sampling an image from a DPM is done by running the reverse process.

[1] For notational brevity, we use the term "marginal" to include distributions conditioned on $\boldsymbol{x}_0$.

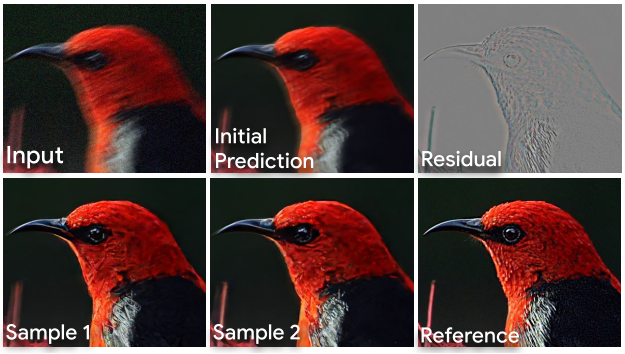

Figure 3. Output of the initial predictor and multiple samples generated from it. We see that the over-smoothed initial prediction lacking texture is "corrected" by the stochastic sampler, producing crisp and diverse final reconstructions. The residual (top right) shows the the difference between reference and initial prediction.

Given some inference-time noise schedule $\bar{\alpha}_{1:T}$, we start from a pure Gaussian noise $\boldsymbol{x}_T \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_d)$ and repeatedly apply the reverse process transition $p_\theta(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t)$ defined in Eq. (4). Notice that this procedure requires a total of T calls to the denoiser network. At the end of this sampling procedure, we are left with a single sample $\boldsymbol{x}_0$.

4. Predict-and-Refine Diffusion Model

One of the main drawbacks of DPM is the computational cost of generating samples, which may require up to thousands of forward passes of the denoiser network due to the iterative denoising procedure. As such, many recent works have explored alternative sampling strategies that reduce the number of sampling steps [29, 33, 37, 57, 61, 69].

We introduce a simple technique that reduces this cost by exploiting the fact that it is often possible to get a cheap initial guess for conditional generative models. Specifically, we augment our conditional diffusion model with a deterministic *initial predictor* (Fig. 2), which provides a data-adaptive candidate for the clean image. Then the denoiser network only needs to model the *residual*.

Letting g_θ denote the initial predictor, the new objective becomes: $L_{\text{Ours}}(\theta) =$

$$\mathbb{E}\left\|\boldsymbol{\epsilon} - f_\theta\Big(\sqrt{\bar{\alpha}}\big(\underbrace{\boldsymbol{x}_0 - g_\theta(\boldsymbol{x}_0)}_{\text{residual}}\big) + \sqrt{1-\bar{\alpha}}\boldsymbol{\epsilon}, \bar{\alpha}, \boldsymbol{y}\Big)\right\|_1 \quad (6)$$

We include a pseudocode for the modified sampling procedure in Algorithm 1. Notice that the initial predictor g_θ does not require an extra loss or pretraining because the gradient from the loss flows through f_θ into g_θ.

Since the initial predictor runs only once, it is beneficial to keep the denoiser network small by offloading most of the computation to the initial predictor. This leads to much more efficient sampling because any reduction in the computational cost of the denoiser network gets amplified by the number of sampling steps used. We further explore this effect in Sec. 6.

Algorithm 1 Predict-and-refine diffusion sampling. The expressions for $\boldsymbol{\mu}_t, \bar{\alpha}_t, \beta_t$ can be found in Sec. 3.

Require: f_θ: Denoiser network, g_θ: Initial predictor, $\boldsymbol{y}$: Blurry input image, $\boldsymbol{\alpha}_{1:T}$: Noise schedule.

$\boldsymbol{x}_{\text{init}} \leftarrow g_\theta(\boldsymbol{y})$ ▷ Initial prediction
$\boldsymbol{z}_T \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_d)$ ▷ Run diffusion sampling
for $t = T, \ldots, 1$ **do**
  $\boldsymbol{\epsilon}_t \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_d)$
  $\boldsymbol{z}_{t-1} \leftarrow \boldsymbol{\mu}_t(\boldsymbol{z}_t, f_\theta(\boldsymbol{z}_t, \bar{\alpha}_t, \boldsymbol{y})) + \beta_t \boldsymbol{\epsilon}_t$
    ▷ Reverse diffusion step; see Eq. (3)
end for
return $\boldsymbol{x}_{\text{init}} + \boldsymbol{z}_0$ ▷ Return the final restoration

4.1. Perception-Distortion Trade-off

As explained in Sec. 3, conditioning the diffusion model on continuous noise level makes it possible to use a different noise schedule during inference. We observe that using many steps with small noise level generally leads to better perceptual quality, and using fewer steps with large noise level leads to lower distortion.

For our experiments, we run a small grid search over the noise schedule hyperparameters and use the model with the best LPIPS score (labeled "Ours"). We emphasize that this inference-time hyperparameter tuning is cheap as it does not involve retraining or finetuning the model itself.

Sample averaging. Our framework also provides a principled alternative to geometric self-ensemble [41]. Since our stochastic sampler is trained to learn the target posterior $p(\boldsymbol{x} \mid \boldsymbol{y})$, we can average multiple samples from our model to approximate the conditional mean $\mathbb{E}[\boldsymbol{x} \mid \boldsymbol{y}]$, *i.e.* the minimum mean squared error estimator. We thus report results for a second model (labeled "Ours-SA") that returns the average of multiple samples.

Traversing the Perception-Distortion curve. By appropriately setting the inference-time hyperparameters mentioned above (sampling steps T, noise schedule $\bar{\alpha}_{1:T}$, and sample averaging), we can smoothly traverse the P-D curve as shown in Fig. 1.

For example, the LPIPS-optimized model ("Ours") uses a relatively large step count of $T = 500$ *without* sample averaging to achieve high perceptual quality at a slight cost of PSNR. The distortion-optimized model ("Ours-SA") does the opposite by using $T = 10$ with sample averaging to sacrifice perceptual quality for higher PSNR. Each point on the P-D curve in Fig. 1 thus corresponds to a specific choice of these hyperparameters.

4.2. Resolution-agnostic Architecture

Unlike the image benchmarks commonly used to evaluate DPMs, blind deblurring benchmarks contain images with various sizes. To support arbitrary input shapes, we use a fully-convolutional architecture for both initial predictor and denoiser network.

Our architecture is based on SR3 [56], which uses a variant of U-Net architecture from [24] with residual blocks replaced with that of BigGAN [7]. To make our model agnostic to image resolution, we removed self-attention, positional encoding, and group normalization. The exact specification of our architecture can be found in the supplementary material.

We note that, to the best of our knowledge, this is the first time a conditional diffusion model is made to support arbitrary image size. Our preliminary experiments show that the fully-convolutioanl architecture had little to no degradation in sample quality for deblurring at non-native resolutions. Because the denoiser network is a relatively simple U-Net, DPMs provide a particularly convenient choice for conditional image generation that must work on any input size.

5. Experiments

5.1. Datasets

We train and evaluate our models on two widely-used image deblurring datasets. For a fair comparison, we follow the same setup used by [11, 14, 35, 48, 63, 74] and train our model only using the provided training data.

GoPro. GoPro dataset [48] contains 3214 pairs of clean and blurry 1280×720 images, of which 1111 are reserved for evaluation. These images are generated by recording video clips with high shutter speed, then averaging consecutive frames to simulate blurs caused by slow shutter speed.

HIDE. We additionally evaluate our GoPro-trained model on the HIDE [59] dataset, which contains 2025 images also of size 1280×720. By training and evaluating our model on different datasets, we can test its ability to generalize under a distributional shift.

5.2. Model Training

We jointly train the initial predictor and denoiser network by minimizing the loss in Eq. (6). Since our model is fully convolutional, we use random 128×128 crops during training, but apply the model on full-size images for evaluation. We also perform training-time data augmentation with random horizontal/vertical flips and $90°/180°/270°$ rotations.

A note on training data. Most currently leading methods only report distortion-based metrics (PSNR and SSIM) and provide pre-trained models for GoPro. Since our work focuses on perceptual quality, we need to compute perceptual metrics ourselves using outputs from other methods. Thus to ensure a fair comparison, we are limited to using models trained on the GoPro dataset, as it is the only dataset with widely available pre-trained models. Nonetheless, we provide additional results and the details of how we obtained the outputs of other methods in the supplementary material.

Table 1. Image deblurring results on the GoPro [48] dataset. Our proposed method sets the new Pareto frontier in terms of Perception-Distortion trade-off. Best values and second-best values for each metric are color-coded. KID values are scaled by a factor of 1000 for readability.

	Perceptual				Distortion	
	LPIPS↓	NIQE↓	FID↓	KID↓	PSNR↑	SSIM↑
Ground Truth	0.0	3.21	0.0	0.0	∞	1.000
HINet [11]	0.088	4.01	17.91	8.15	32.77	0.960
MPRNet [73]	0.089	4.09	20.18	9.10	32.66	0.959
MIMO-UNet+ [14]	0.091	4.03	18.05	8.17	32.45	0.957
SAPHNet [63]	0.101	3.99	19.06	8.48	31.89	0.953
SimpleNet [40]	0.108				31.52	0.950
DeblurGANv2 [35]	0.117	3.68	13.40	4.41	29.08	0.918
Ours	0.059	3.39	4.04	0.98	31.66	0.948
Ours-SA	0.078	4.07	17.46	8.03	33.23	0.963

5.3. Evaluation

Evaluation Metrics. We evaluate our method on four different perceptual metrics: LPIPS [75], NIQE [47], FID (Fréchet Inception Distance) [23], and KID (Kernel Inception Distance) [4]. Because our datasets do not have enough examples to reliably compute FID and KID, we extract 15 non-overlapping patches of size 256×240 from each 1280×720 image and compute the Inception-based metrics at the patch level, similar to [46]. For completeness, we also include two distortion-based metrics: PSNR and SSIM [68].

We note the importance of including full-reference metrics for conditional image generation. A method can achieve near-perfect score on a no-reference metric such as NIQE by producing highly realistic images that are completely unrelated to the input. This is particularly relevant for GAN-based methods, since the discriminator may not penalize the generator for producing natural-looking images that do not match the input. This is why we included LPIPS (and to some extent, PSNR and SSIM), even though it is technically not a perceptual metric. For a qualitative comparison, we also conduct a human study and provide sample restorations.

5.4. Quantitative Results

5.4.1 GoPro Results

Table 1 shows quantitative results on the GoPro dataset. We compared our model with the current state-of-the-art (SOTA) methods HINet [11], MPRNet [73], and DeblurGAN-v2 [35].

Figure 4. Sample deblurred images from GoPro and HIDE datasets. Because our method is not trained to minimize distortion-based loss (*e.g.* L_2), it avoids producing blurry output and achieves better reconstruction of detailed textures. Full-size images are provided in the supplementary material. Best viewed electronically.

Our model achieves SOTA performance across **all perceptual metrics** while maintaining competitive PSNR and SSIM to existing methods. Notably, we obtain the FID of 4.04, **nearly a 70% reduction** compared to DeblurGAN-v2 [35], the current SOTA method in terms of perceptual quality. Moreover, the sample-averaging variant of our method achieves a new SOTA PSNR of 33.23 while still outperforming all other methods with respect to LPIPS. All in all, these results highlight our framework's flexibility to control the trade-off between perception and distortion using a single model. As shown in Figure 1, our result sets a new Pareto frontier on the Perception-Distortion plot.

Table 2. Image deblurring results on the HIDE [59] dataset, using models trained on GoPro [48]. Our method significantly outperforms the baseline methods under all perceptual metrics while maintaining competitive PSNR and SSIM. Best values and second-best values for each each metric are color-coded.

	Perceptual				Distortion	
	LPIPS↓	NIQE↓	FID↓	KID↓	PSNR↑	SSIM↑
Ground Truth	0.0	2.72	0.0	0.0	∞	1.000
HINet [11]	0.120	3.20	15.17	7.33	30.33	0.932
MIMO-UNet+ [14]	0.124	3.24	16.01	7.91	29.99	0.930
MPRNet [73]	0.114	3.46	16.58	8.35	30.96	0.939
SAPHNet [63]	0.128	3.21	16.77	8.39	29.99	0.930
DeblurGAN-v2 [35]	0.159	2.96	15.51	6.97	27.51	0.885
Ours	0.089	2.69	5.43	1.61	29.77	0.922
Ours-SA	0.092	2.93	6.37	2.40	30.07	0.928

5.4.2 HIDE Results

We also evaluate our GoPro-trained model on the HIDE dataset [59] to test its ability to generalize to out-of-distribution input. As the results in Table 2 clearly show, the gains in perceptual quality do translate over to the HIDE dataset. In particular, both of our models significantly outperform the baseline methods across **all perceptual metrics** while maintaining competitive distortion values.

Fig. 4 includes several sample reconstructions from both GoPro and HIDE datasets. Despite sometimes containing a little more noise (some of which was presumably learned from the training data itself), we see that our model shows a clear improvement in perceptual quality. Additional full-size comparisons are provided in the supplementary material.

5.5. Human Study for Qualitative Evaluation

We ran a perceptual study with human subjects to further quantify the performance of the proposed deblurring framework. Our results are presented in Table 3. We used Amazon Mechanical Turk to obtain pairwise ratings comparing different deblurring methods applied on the GoPro dataset. In this study, the human subjects had a minimum of 70% approval rating, and were asked to select the image with the better quality from side-by-side crops of size 512×512.

Results in Table 3 show the average rater's preference computed from 480 comparisons. As the highlighted cells show, these results indicate that **both variations of our deblurring model outperform the competing methods**.

We also observed that raters showed a modest preference for the sample-averaged variant in crops with relatively flat content. On the other hand, raters preferred individual samples for highly-textured crops. Fig. 5 shows that the level of detail produced by our model is adaptive to the blur present in the input. As expected, blurrier images generally lead to higher variance in the resulting samples.

Table 3. Average pairwise human preference for deblurring results on the GoPro dataset [48]. Each value represents the percentage of times Amazon Mechanical Turk raters chose the row over the column. Each preference percentage is an average over 480 ratings (20 raters, and 24 unique image pairs).

	HINet	MPRNet	Ours	Ours-SA	Reference
HINet [11]	-	54.9	29.1	31.0	14.5
MPRNet [73]	45.1	-	26.6	25.3	11.9
Ours	70.9	73.4	-	58.8	37.1
Ours-SA	69.0	74.7	41.2	-	26.7
Reference	85.5	88.1	62.9	73.3	-

Figure 5. Deblurred samples for crops of two different images. The ill-posedness of the restoration task (*i.e.* strength of the blur) has a direct impact on the diversity of the generated samples. This is illustrated by the per-pixel standard deviation computed using multiple restorations for each input image. As clearly visible in the right-most column, the blurrier input (first row) corresponds to overall higher per-pixel standard deviations.

6. Discussion and Analysis

For the analysis of various aspects of our model, we used a custom dataset created by applying synthetic camera shake blur and noise (described in the supplementary material) on the images of the DIV2K dataset [1]. This was done to make qualitative evaluation in a more controlled environment, since the low-quality ground truth images in existing paired datasets [48, 54] make qualitative assessment difficult.

6.1. Benefits of Residual Modeling

More efficient sampling. The main benefit of residual modeling is the reduction in the computational cost of sampling. Due to the iterative nature of diffusion sampling, the denoiser network must run many times for each generated sample – sometimes up to hundreds to thousands of steps. Thus, any reduction in the cost of running the denoiser is particularly valuable, and our initial predictor provides a simple way to offload some of this computation.

A key question is then whether the initial predictor can compensate for the decrease in the sample quality from using a smaller denoiser network. We empirically explore this by comparing sampling latency against sample quality with and without the initial predictor. In Fig. 6, the non-residual

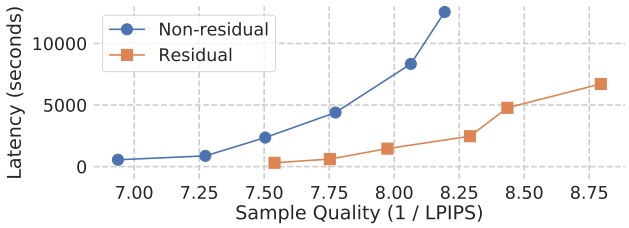

Figure 6. Plot of sampling cost vs. sample quality. Even with the added parameters from the initial predictor, the residual model achieves lower latency while maintaining higher sample quality.

model refers to a regular conditional diffusion model with a large denoiser network. The residual model follows our architecture and has a large initial predictor and a small denoiser. Overall, the residual model has more parameters (33M vs. 28M).

We see that the residual model requires much less time to sample an image despite it being larger than the non-residual model. Importantly, this reduction in sampling cost does not negatively affect the sample quality – in fact, the residual model is up to 7× faster for a comparable sample quality.

Output of the initial predictor. One unexpected discovery from our experiments is that the output of the initial predictor is often a fairly reasonable reconstruction of the reference image. We can see this in Fig. 3. While lacking in detail, the initial prediction is certainly less blurry than the input.

It is perhaps surprising that this happens even though there is no explicit loss on the initial predictor's output $g_\theta(\boldsymbol{y})$ to match the reference. We also note that our method is not the only possible parameterization of a diffusion model with an explicit decoupling of the iterative portion (denoiser network) from the single-pass portion (initial predictor). For instance, we could have simply fed $g_\theta(\boldsymbol{y})$ as an auxiliary input to the denoiser f_θ without computing the residual. We leave these investigations around the initial predictor as future work.

Residual images are simpler to model. One may wonder why adding a deterministic initial predictor would help with the model's performance. We posit that the benefits of residual modeling may be due to the distribution of residual images being "simpler" than that of reference images.

While it is impractical to approximate the true entropy of the two distributions, we can look at related quantities that may serve as a proxy. Specifically, we compute the entropy of pixel values aggregated across all pixel locations for residual and reference images. As expected from natural images, the reference pixel distribution is reasonably spread out and has the entropy of 7.42 bits-per-dimension (bpd). On the other hand, the residual pixel values follow a much more sharply concentrated distribution, leading to a substantially lower entropy of 3.91 bpd. This suggests that the residual images may indeed be simpler to model.

6.2. Network Architecture Ablation

To better understand where the performance gains of our method are originating from, we trained a regression-based baseline that only uses the initial predictor. Surprisingly, we observed that the initial predictor alone was able to achieve state-of-the-art PSNR of 33.07 when trained with a simple L_2 loss. Through a detailed ablation study, we identified three key hyperparameters: exponential moving average (EMA) of weights, large batch size, and network size.

In Table 4, we start from a simple U-Net architecture [55] and gradually enable each of the aforementioned hyperparameters. All models were trained for 1M steps to ensure the differences are not due to insufficient training. As the results show, all three hyperparameters were critical to the model's performance.

Table 4. Ablation study on the effects of various hyperparameters for our U-Net architecture, evaluated on the GoPro dataset.

	Hyperparameters			Metrics			
	ch.	batch	EMA	LPIPS	PSNR	MParam.	BFLOPs
More Channels	16	32	No	0.137	29.93	1.63	301
	32	32	No	0.113	31.05	6.52	1200
	64	32	No	0.103	31.63	26.07	4790
+Larger Batch	64	64	No	0.099	31.85	26.07	4790
	64	128	No	0.087	32.56	26.07	4790
	64	256	No	0.086	32.61	26.07	4790
+Use EMA	64	256	Yes	0.0809	33.07	26.07	4790

7. Conclusion and Future Directions

We presented a new framework for stochastic blind image deblurring with a focus on perceptual quality using a conditional diffusion model. We introduced a novel technique for reducing the computational burden of diffusion sampling. We empirically showed that our method achieves significantly improved perceptual quality and competitive distortion metrics as compared to the current state-of-the-art methods. We believe that our work opens a new direction for blind deblurring with a focus on perceptual quality and establishes a strong benchmark for future works to improve upon.

There are a number of avenues to explore to further address the limitations of our work. Due to slow sampling and large network size, diffusion models are computationally too expensive to be incorporated into consumer-level devices. One way to combat this is to use more efficient sampling schemes such as DDIM [61] or distillation [2]. Another promising direction is to replace our initial predictor and denoiser network with U-Net architectures that are optimized for both distortion and run time [11, 14, 73].

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*, July 2017. 7

[2] Anonymous. Progressive distillation for fast sampling of diffusion models. In *Submitted to The Tenth International Conference on Learning Representations*, 2022. under review. 8

[3] Muhammad Asim, Fahad Shamshad, and Ali Ahmed. Blind image deconvolution using deep generative priors. *IEEE Transactions on Computational Imaging*, 6:1493–1506, 2020. 2

[4] Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans. In *International Conference on Learning Representations*, 2018. 1, 5

[5] 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. 1, 2

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

[7] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. In *International Conference on Learning Representations*, 2018. 5

[8] Ayan Chakrabarti. A neural approach to blind motion deblurring. In *European conference on computer vision*, pages 221–235. Springer, 2016. 2

[9] Tony F Chan and Chiu-Kwong Wong. Total variation blind deconvolution. *IEEE transactions on Image Processing*, 7(3):370–375, 1998. 1, 2

[10] Liang Chen, Faming Fang, Tingting Wang, and Guixu Zhang. Blind image deblurring with local maximum gradient prior. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1742–1750, 2019. 2

[11] Liangyu Chen, Xin Lu, Jie Zhang, Xiaojie Chu, and Chengpeng Chen. Hinet: Half instance normalization network for image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 182–192, June 2021. 2, 5, 7, 8

[12] Nanxin Chen, Yu Zhang, Heiga Zen, Ron J Weiss, Mohammad Norouzi, and William Chan. Wavegrad: Estimating gradients for waveform generation. In *International Conference on Learning Representations*, 2020. 3

[13] Nanxin Chen, Yu Zhang, Heiga Zen, Ron J Weiss, Mohammad Norouzi, Najim Dehak, and William Chan. Wavegrad 2: Iterative refinement for text-to-speech synthesis. *arXiv preprint arXiv:2106.09660*, 2021. 3

[14] Sung-Jin Cho, Seo-Won Ji, Jun-Pyo Hong, Seung-Won Jung, and Sung-Jea Ko. Rethinking coarse-to-fine approach in single image deblurring. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4641–4650, October 2021. 1, 2, 5, 7, 8

[15] 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. 3

[16] Mauricio Delbracio, Ignacio Garcia-Dorado, Sungjoon Choi, Damien Kelly, and Peyman Milanfar. Polyblur: Removing mild blur by polynomial reblurring. *IEEE Transactions on Computational Imaging*, 7:837–848, 2021. 2

[17] M. Delbracio, H. Talebei, and P. Milanfar. Projected distribution loss for image enhancement. In *2021 IEEE International Conference on Computational Photography (ICCP)*, pages 1–12. IEEE Computer Society, 2021. 2

[18] Rob Fergus, Barun Singh, Aaron Hertzmann, Sam T Roweis, and William T Freeman. Removing camera shake from a single photograph. In *ACM SIGGRAPH 2006 Papers*, pages 787–794. 2006. 1, 2

[19] Dror Freirich, Tomer Michaeli, and Ron Meir. A theory of the distortion-perception tradeoff in wasserstein space. *arXiv preprint arXiv:2107.02555*, 2021. 2

[20] Hongyun Gao, Xin Tao, Xiaoyong Shen, and Jiaya Jia. Dynamic scene deblurring with parameter selective sharing and nested skip connections. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3848–3856, 2019. 2

[21] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2414–2423, 2016. 2

[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. 2

[23] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. 5

[24] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851. Curran Associates, Inc., 2020. 3, 5

[25] Ajil Jalal, Marius Arvinte, Giannis Daras, Eric Price, Alexandros G Dimakis, and Jonathan I Tamir. Robust compressed sensing mri with deep generative priors. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 2

[26] Ajil Jalal, Sushrut Karmalkar, Jessica Hoffmann, Alex Dimakis, and Eric Price. Fairness for image generation with uncertain sensitive attributes. In *International Conference on Machine Learning*, pages 4721–4732. PMLR, 2021. 3

[27] Ajil Jalal, Liu Liu, Alexandros G Dimakis, and Constantine Caramanis. Robust compressed sensing using generative models. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 713–727. Curran Associates, Inc., 2020. 3

[28] Meiguang Jin, Stefan Roth, and Paolo Favaro. Normalized blind deconvolution. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 668–684, 2018. 1, 2

[29] Alexia Jolicoeur-Martineau, Ke Li, Rémi Piché-Taillefer, Tal Kachman, and Ioannis Mitliagkas. Gotta go fast when generating data with score-based models. *arXiv preprint arXiv:2105.14080*, 2021. 4

[30] Zahra Kadkhodaie and Eero P Simoncelli. Stochastic solutions for linear inverse problems using the prior implicit in a denoiser. In *Thirty-Fifth Conference on Neural Information Processing Systems*, 2021. 3

[31] Bahjat Kawar, Gregory Vaksman, and Michael Elad. SNIPS: Solving noisy inverse problems stochastically. In *Thirty-Fifth Conference on Neural Information Processing Systems*, 2021. 3

[32] Bahjat Kawar, Gregory Vaksman, and Michael Elad. Stochastic image denoising by sampling from the posterior distribution. In *Proceedings of the International Conference on Computer Vision (ICCV) Workshops*, 2021. 3

[33] Zhifeng Kong and Wei Ping. On fast sampling of diffusion probabilistic models. In *ICML Workshop on Invertible Neural Networks, Normalizing Flows, and Explicit Likelihood Models*, 2021. 4

[34] 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. 2, 3

[35] 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)*, Oct 2019. 1, 2, 3, 5, 6, 7

[36] Wei-Sheng Lai, Jia-Bin Huang, Zhe Hu, Narendra Ahuja, and Ming-Hsuan Yang. A comparative study for single image blind deblurring. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1701–1709, 2016. 2

[37] Sang-gil Lee, Heeseung Kim, Chaehun Shin, Xu Tan, Chang Liu, Qi Meng, Tao Qin, Wei Chen, Sungroh Yoon, and Tie-Yan Liu. Priorgrad: Improving conditional denoising diffusion models with data-driven adaptive prior. *arXiv preprint arXiv:2106.06406*, 2021. 4

[38] Anat Levin, Yair Weiss, Fredo Durand, and William T Freeman. Understanding and evaluating blind deconvolution algorithms. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1964–1971. IEEE, 2009. 1, 2

[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. 3

[40] Jichun Li, Weimin Tan, and Bo Yan. Perceptual variousness motion deblurring with light global context refinement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4116–4125, October 2021. 1, 5

[41] 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. 4

[42] 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. 2, 3

[43] 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. 3

[44] Roey Mechrez, Itamar Talmi, Firas Shama, and Lihi Zelnik-Manor. Maintaining natural image statistics with the contextual loss. In *Asian Conference on Computer Vision*, pages 427–443. Springer, 2018. 2

[45] Roey Mechrez, Itamar Talmi, and Lihi Zelnik-Manor. The contextual loss for image transformation with non-aligned data. In *European Conference on Computer Vision (ECCV)*, pages 768–783, 2018. 2

[46] Fabian Mentzer, George D Toderici, Michael Tschannen, and Eirikur Agustsson. High-fidelity generative image compression. *Advances in Neural Information Processing Systems*, 33, 2020. 5

[47] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal processing letters*, 20(3):209–212, 2012. 5

[48] 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)*, July 2017. 2, 5, 7

[49] Guy Ohayon, Theo Adrai, Gregory Vaksman, Michael Elad, and Peyman Milanfar. High perceptual quality image denoising with a posterior sampling cgan. In *Proceedings of the International Conference on Computer Vision (ICCV) Workshops*, 2021. 3

[50] Gregory Ongie, Ajil Jalal, Christopher A Metzler, Richard G Baraniuk, Alexandros G Dimakis, and Rebecca Willett. Deep learning techniques for inverse problems in imaging. *IEEE Journal on Selected Areas in Information Theory*, 1(1):39–56, 2020. 2

[51] Mangal Prakash, Alexander Krull, and Florian Jug. Fully unsupervised diversity denoising with convolutional variational autoencoders. In *International Conference on Learning Representations*, 2020. 3

[52] Sainandan Ramakrishnan, Shubham Pachori, Aalok Gangopadhyay, and Shanmuganathan Raman. Deep generative filter for motion deblurring. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 2993–3000, 2017. 2

[53] Wenqi Ren, Jiawei Zhang, Jinshan Pan, Sifei Liu, Jimmy Ren, Junping Du, Xiaochun Cao, and Ming-Hsuan Yang. Deblurring dynamic scenes via spatially varying recurrent neural networks. *IEEE transactions on pattern analysis and machine intelligence*, 2021. 2

[54] Jaesung Rim, Haeyun Lee, Jucheol Won, and Sunghyun Cho. Real-world blur dataset for learning and benchmarking deblurring algorithms. In *European Conference on Computer Vision*, pages 184–201. Springer, 2020. 7

[55] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 2, 8

[56] 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. 3, 5

[57] Robin San-Roman, Eliya Nachmani, and Lior Wolf. Noise estimation for generative diffusion models. *arXiv preprint arXiv:2104.02600*, 2021. 4

[58] Qi Shan, Jiaya Jia, and Aseem Agarwala. High-quality motion deblurring from a single image. *Acm transactions on graphics (tog)*, 27(3):1–10, 2008. 1, 2

[59] 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. 1, 5, 7

[60] 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*, volume 37 of *Proceedings of Machine Learning Research*, pages 2256–2265. PMLR, 07–09 Jul 2015. 3

[61] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021. 4, 8

[62] Shuochen Su, Mauricio Delbracio, Jue Wang, Guillermo Sapiro, Wolfgang Heidrich, and Oliver Wang. Deep video deblurring for hand-held cameras. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1279–1288, 2017. 2

[63] Maitreya Suin, Kuldeep Purohit, and A. N. Rajagopalan. Spatially-attentive patch-hierarchical network for adaptive motion deblurring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 1, 5, 7

[64] Jian Sun, Wenfei Cao, Zongben Xu, and Jean Ponce. Learning a convolutional neural network for non-uniform motion blur removal. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 769–777, 2015. 2

[65] Xin Tao, Hongyun Gao, Xiaoyong Shen, Jue Wang, and Jiaya Jia. Scale-recurrent network for deep image deblurring. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 1, 2

[66] Fu-Jen Tsai, Yan-Tsung Peng, Yen-Yu Lin, Chung-Chi Tsai, and Chia-Wen Lin. Banet: Blur-aware attention networks for dynamic scene deblurring. *arXiv preprint arXiv:2101.07518*, 2021. 1

[67] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022*, 2016. 2

[68] 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. 2, 5

[69] Daniel Watson, Jonathan Ho, Mohammad Norouzi, and William Chan. Learning to efficiently sample from diffusion probabilistic models. *arXiv preprint arXiv:2106.03802*, 2021. 4

[70] Jay Whang, Erik Lindgren, and Alex Dimakis. Composing normalizing flows for inverse problems. In *International Conference on Machine Learning*, pages 11158–11169. PMLR, 2021. 3

[71] Patrick Wieschollek, Michael Hirsch, Bernhard Scholkopf, and Hendrik Lensch. Learning blind motion deblurring. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 231–240, 2017. 2

[72] Li Xu, Shicheng Zheng, and Jiaya Jia. Unnatural l0 sparse representation for natural image deblurring. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1107–1114, 2013. 2

[73] 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, June 2021. 2, 5, 7, 8

[74] Hongguang Zhang, Yuchao Dai, Hongdong Li, and Piotr Koniusz. Deep stacked hierarchical multi-patch network for image deblurring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 1, 5

[75] 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)*, June 2018. 5

[76] Xiang Zhu, Filip Šroubek, and Peyman Milanfar. Deconvolving psfs for a better motion deblurring using multiple images. In *European Conference on Computer Vision*, pages 636–647. Springer, 2012. 2