Multiscale Structure Guided Diffusion for Image Deblurring

Mengwei Ren[†‡*] Mauricio Delbracio[‡] Hossein Talebi[‡] Guido Gerig[†] Peyman Milanfar[‡]
[†]New York University [‡]Google Research

Abstract

Diffusion Probabilistic Models (DPMs) have recently been employed for image deblurring, formulated as an image-conditioned generation process that maps Gaussian noise to the high-quality image, conditioned on the blurry input. Image-conditioned DPMs (icDPMs) have shown more realistic results than regression-based methods when trained on pairwise in-domain data. However, their robustness in restoring images is unclear when presented with out-of-domain images as they do not impose specific degradation models or intermediate constraints. To this end, we introduce a simple yet effective multiscale structure guidance as an implicit bias that informs the icDPM about the coarse structure of the sharp image at the intermediate layers. This guided formulation leads to a significant improvement of the deblurring results, particularly on unseen domain. The guidance is extracted from the latent space of a regression network trained to predict the clean-sharp target at multiple lower resolutions, thus maintaining the most salient sharp structures. With both the blurry input and multiscale guidance, the icDPM model can better understand the blur and recover the clean image. We evaluate a single-dataset trained model on diverse datasets and demonstrate more robust deblurring results with fewer artifacts on unseen data. Our method outperforms existing baselines, achieving state-of-the-art perceptual quality while keeping competitive distortion metrics.

1. Introduction

Image deblurring is a fundamentally ill-posed inverse problem that aims to estimate one (or several) high-quality image(s) given a blurry observation. Deep networks allow for end-to-end image deblurring with pairwise supervised learning. While deep regression-based methods [81, 90, 97, 79, 6, 88, 7, 83, 78, 41, 23, 52] optimize distortion metrics such as PSNR, they often produce over-smoothed outputs that lack visual fidelity [39, 5, 13, 4]. Therefore, perceptual-driven methods [43, 26] aim to produce sharp and visually pleasing images that are still faithful to the sharp reference image, typically with a slight compromise on distortion performance, i.e., a less than 3dB drop on PSNR [4, 55] allows for significantly better visual quality while still being close to the target image. GANs [16] are leveraged for improved deblurring perception [36, 37]. However, GAN training suffers from instability, mode-collapse and artifacts [47], which may hamper the plausibility of the generated images.

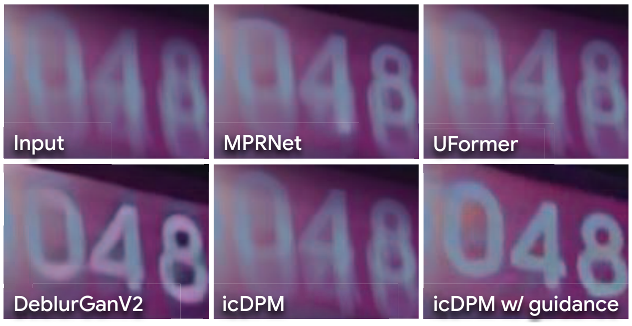

Figure 1. Deblurring example on Realblur-J dataset [60] with models *only* trained on synthetic GoPro data [51], from recent regression-based [88, 82] (MPRNet, UFormer), GAN-based [37] (DeblurGANV2) and image-conditioned diffusion probabilistic methods (icDPM). We introduce a guidance module onto the icDPM formulation, and improves its robustness on unseen image.

Recently, DPMs [18] further improved the photo-realism in a variety of imaging inverse problems [67, 42, 83, 65, 12], formulated as an image-conditioned generation process, where the DPM takes the degraded estimation as an auxiliary input. Image-conditioned DPMs (icDPMs) do not estimate the degradation kernel nor impose any intermediate constraints. These models are trained using a standard denoising loss [18] with pairwise training data in a supervised fashion. In image restoration, such pairwise training dataset is typically artificially curated by applying known degradation models on a group of clean images, which inevitably introduces a domain gap between the synthetic training dataset and real-world blurry images. When presented with unseen data, the robustness of icDPMs are rather unclear as the intermediate restoration process is intractable. E.g., we observe a noticeable performance drop when we apply the synthetically trained icDPM to out-of-domain data,

[1] Work done during an internship at Google Research.

including failure to deblur the input (Fig. 1) and injection of artifacts (Fig. 4 'icDPM' and Fig. 7 'DvSR'). We empirically established a connection between domain sensitivity and image-conditioning in the existing deblurring icDPMs [65, 67, 83], where the observed poor generalization is attributed to the naive input-level concatenation and the lack of intermediate constraints during the deblurring process. When optimized on the synthetic training set, overfitting or memorization [71] may occur, making the model vulnerable to shift of the input distribution. Currently, conditioning DPM on blurred or corrupted images is under-explored [61], and we hypothesize that more effective image conditioning for icDPM is crucial to make the model more constrained and robust towards unseen domain.

Inspired by traditional blind deblurring algorithms where optimization is made using explicit structural priors (e.g., containing image saliency [56, 85]), we enhance the icDPM backbone (UNet [63]) with a multiscale structure guidance at intermediate layers. These guidance features are obtained through a regression network trained to predict salient sharp features from the input. The guidance, in conjunction with the blurry image, provide more informative cues to the model regarding the specific degradation in the image. As a result, the model can more accurately recover the clean image and generalize more effectively. Our contributions are threefold: (1) we investigate and analyze the domain generalization of conditional diffusion models in motion deblurring task, and empirically find a relationship between model robustness and image-conditioning; (2) we propose an intuitive but effective guidance module that projects the input image to a multiscale structure representation, which is then incorporated as an auxiliary prior to make the diffusion model more robust; (3) Compared with existing benchmarks, our single-dataset trained model shows more robust results across different test sets by producing more plausible deblurring and fewer artifacts, quantified by the state-of-the-art perceptual quality and on par distortion metrics.

2. Related Works

Single image deblurring is the inverse process of recovering one or multiple high-quality, sharp images from the blurry observation. Typically, classic deblurring approaches involve variational optimization [15, 35, 40, 48, 57, 85, 1, 25], with prior assumptions on blur kernels, images or both, to alleviate the ill-posedness of the inverse problem. Handcrafted structural priors, such as edges and shapes, have been used successfully in many algorithms to guide the deblurring process towards preserving important features in the image while removing blur [56, 57, 85]. Our design principle is inspired by these approaches and involves a learned guidance as implicit structural bias. With the emergent of deep learning, deblurring can be cast as a particular image-to-image translation problem where a deep model takes the blurry image as its input, and predicts a high quality counterpart, supervised by pixel wise losses between the recovered image and the target [81, 90, 97, 79, 6, 88, 7, 83, 78, 41, 23, 52, 24]. Pixel-wise losses, such as L_1 and L_2, are known to result in over-smoothed images [39, 5, 13] given their 'regression to the mean' nature. To this end, perception-driven losses including perceptual [26, 93, 46, 45, 95, 13] and adversarial losses [36, 37] are added on top of the pixel-wise constraints, to improve the visual fidelity of the deblurred image, with a compromising drop in distortion scores [4, 55]. Tangentially, recent works seek to improve the architectural design by exploring attention mechanisms [54, 87, 82, 86, 78, 79], multi-scale paradigms [51, 7] and multi-stage frameworks [89, 6, 88].

Diffusion Probablistic Models (DPM) [70, 18, 72, 14], Score-based models [74, 75, 76] and their recent exploratory generalizations [2, 22, 11] achieved remarkable results in a varied range of applications[9], from image and video synthesis[66, 58, 62, 19, 30, 20], to solving general imaging inverse problems[10, 27, 32, 29, 38, 8]. DPMs are characterized for having stable training [18, 14, 28], diverse mode coverage [73, 33], and high perception [66, 14, 58]. DPM formulation involves a fixed forward process of gradually adding Gaussian noise to the image, and a learnable reverse process to denoise and recover the clean image, operated with a Markov chain structure. Conditional DPMs aim to perform image synthesis with an additional input (class [14], text [66, 58], source image).

Image-conditioned DPMs (icDPMs) have been successfully re-purposed for image restoration tasks such as super-resolution [67, 42], deblurring [83], JPEG restoration [65, 31]. This is achieved by concatenating the corrupted observation at input level. They do not require task-specific losses or architectural designs, and have been adopted due to high sample perceptual quality. ControlNet [92] further enables the task-specific image conditioning for pretrained text-to-image diffusion models. InDI [12] presents an alternative, and more intuitive, diffusion process where the low-quality input is directly restored into a high-quality image in small steps. Nevertheless, generalization of DPMs to unseen shifts in domain, and their low-quality/corrupted image conditioning remains unexplored.

Generalization to unseen domain As mentioned above, deep restoration models for deblurring rely on synthetic pairwise training data. However, any well-trained deep restoration model may fail to produce comparable results on out-of-domain data (Fig. 1). To address this issue, researchers have pursued two main directions for improving model generalization: enhancing the representativeness and realism of the training data, or improving the model's domain generalization ability. Our method focuses on the latter, but it is not mutually exclusive with the former direction and can be combined to further improve the re-

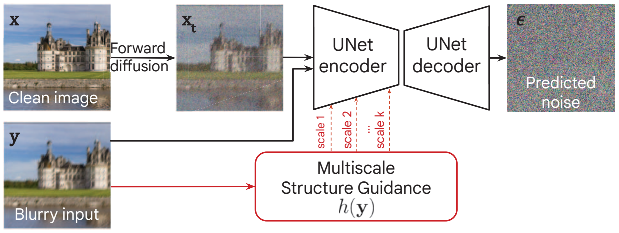

Figure 2. The *training* process of the proposed deblurring method. The backbone model is a standard image-conditioned DPM (icDPM) where a UNet learns to perform denoising towards the clean image conditioned on the blurry input. We equip the icDPM with a structure guidance module (detailed in Fig. 3) to better inform the model of the coarse sharp structure at multiple scales.

sults. To tackle the data limitations, previous works focus on acquiring or combining more representative training data [50, 60, 23, 96], and/or generate realistic degraded images using generative approaches [91, 84]. Other prior works focus on explicit domain adaptation, leveraging transfer learning techniques to reduce domain gaps. These approaches include unpaired image translation [21, 59] and domain adaptation [80, 68, 44, 52], which typically involve an adversarial formulation and joint training between two specified domains. However, these methods may require retraining when a new dataset is introduced. In contrast, our method does not involve explicit adaptation between specified domains. Instead, we focus on introducing more effective image conditioning mechanism that naturally makes the model more robust towards distribution shift.

3. Method

3.1. Overview

We assume access to a paired dataset with samples $(\boldsymbol{x}, \boldsymbol{y}) \sim p_{\text{train}}(\boldsymbol{x}, \boldsymbol{y})$, where $\boldsymbol{x}$ represents the high-quality sharp image, and $\boldsymbol{y}$ is the respective low-quality blurry observation (denoted in Fig. 2). Such paired dataset is typically generated by simulating degraded images from the high-quality ones adopting a specific degradation model. The goal is to reconstruct one or multiple clean, sharp images $\boldsymbol{x}$ from the low-quality observation $\hat{\boldsymbol{y}} \sim p_{\text{real}}(\hat{\boldsymbol{y}})$. Generally, the distribution of the training set p_{train} differs from that of unseen images p_{real}. It is thus crucial that a model not only performs well on p_{train} but also generalizes to p_{real}.

DPMs We consider a general-purpose DPM for our formulation given its superior performance in high-quality image restoration [67, 83]. In what follows, we briefly describe the training and sampling of a DPM to contextualize our work. Unconditional DPMs aim to sample from the data distribution $p(\boldsymbol{x})$ by iteratively denoising samples from a Gaussian distribution and converting them into samples from the target data distribution. To train such model, a forward diffusion process and a reverse process are involved. As illustrated in Fig. 2, at a diffusion step t, a noisy version $\boldsymbol{x}_t$ of the target image $\boldsymbol{x}$ is generated by $\boldsymbol{x}_t = \sqrt{\alpha_t}\boldsymbol{x} + \sqrt{(1-\alpha_t)}\boldsymbol{\epsilon}$, $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \boldsymbol{I_d})$, where $\boldsymbol{\epsilon}$ is sampled from a standard Gaussian distribution $\mathcal{N}(0, \boldsymbol{I}_d)$, and α_t controls the amount of noise added at each step t. In the reverse process, an image-to-image network (i.e. UNet) $\mathcal{G}_\theta(\boldsymbol{x}_t, t)$ parameterized by θ learns to estimate the clean image from the partially noisy input $\boldsymbol{x}_t$. In practice, a reparameterization of the model to predict the noise instead of the clean image leads to better sample quality [18]. Once trained, it samples a clean image by iteratively running for T steps starting from a pure Gaussian noise $\boldsymbol{x}_T \sim \mathcal{N}(0, \boldsymbol{I_d})$.

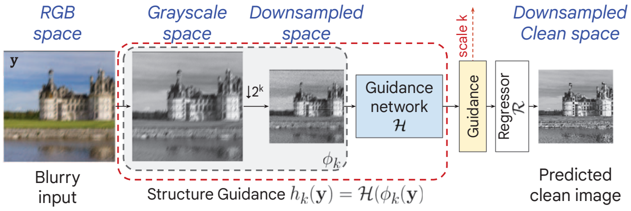

Figure 3. The learned structure guidance (red box in Fig. 2) is extracted from the latent features of a regression network trained to predict the luma channel of the sharp target at multiple lower resolutions. In this way, the guidance maintains only structure-relevant information, representing the underlying sharp image.

Image-conditioned DPMs further inject an input image $\boldsymbol{y}$ so as to generate high-quality samples that are paired with the low-quality observation. This involves generating samples from the conditional distribution of $p(\boldsymbol{x}|\boldsymbol{y})$ (posterior). A conditional DPM $\mathcal{G}_\theta([\boldsymbol{x}_t, \boldsymbol{y}], t)$ is used, where the image conditioning is typically implemented via concatenation of $\boldsymbol{y}$ and $\boldsymbol{x}_t$ at input-level [67, 83, 65]. However, we found that this formulation is sensitive to domain shift in input images, and leads to poor generalization ('DPM' in Fig. 1). Moreover, in many cases it introduces visual artifacts ('DvSR' in Fig. 7). We speculate that this is due to the naive image-conditioning (input-level concatenation), which lacks constraints in the intermediate process. Therefore, we integrate a multiscale structure guidance $h(\boldsymbol{y})$ into the latent space of the icDPM backbone, to inform the model about salient image features, such as significant coarse structures that are essential for reconstructing a high-quality image, while disentangling irrelevant information, such as the footprint of blur kernels and color information. To obtain such guidance with the aforementioned characteristics, we propose an auxiliary regression network and leverage its learned features as the realization of the guidance, described below in Sec. 3.2.

3.2. Multiscale structure guidance

Fig. 3 shows the details of our proposed guidance, denoted as $h(\cdot)$. The DPM equipped with such multiscale guidance is better aware of the underlying salient structures of the input, thus it learns to better sample from the target conditional distribution to $p_{\text{real}}(\boldsymbol{x}|\boldsymbol{y})$. Moreover, the distribution of $h(\boldsymbol{y})$ does not change significantly when the input domain changes, so that it can reliably provide the auxiliary structure guidance even when applied to unseen domain. To both ends, we construct the guidance module as $h_k(\cdot) = \mathcal{H}(\phi_k(\cdot))$. At scale k, it consists of an image transformation function $\phi_k(\cdot)$, followed by a regression-driven guidance network $\mathcal{H}$. Specifically, $\phi_k(\cdot)$ transforms the input image $\boldsymbol{y}$ to suppress information not relevant to the coarse sharp image structures (e.g., color, and information about the domain-specific degradation). This ensures that $\mathcal{H}$ operates on a less input-domain-sensitive space. We first convert $\boldsymbol{y}$ to the grayscale space $\bar{\boldsymbol{y}}$, and then, $\bar{\boldsymbol{y}}$ gets downsampled by a factor of 2^k, where $k = 1, 2, 3$. This removes fine details (including the footprint of blur to certain amount), while preserving coarse structures at multiple lower resolutions. Motivated by [84], we also add a small amount of Gaussian noise to mask other domain specific degradations/characteristics, and make the output less sensitive to input domain shift. In recent diffusion models, the addition of noise becomes a common practice [64, 66] as it enhances robustness when dealing with out-of-domain data. Thus,

$$\phi_k(\boldsymbol{y}) = d_{\downarrow k}(\bar{\boldsymbol{y}}) + \boldsymbol{n}, \quad \boldsymbol{n} \sim \mathcal{N}(0, \sigma^2 \boldsymbol{I}). \tag{1}$$

Then, the guidance network $\mathcal{H}_\varphi$ extracts the guidance feature by mapping $\phi_k(\boldsymbol{y})$ onto the representation/latent space as $h_k(\boldsymbol{y}) = \mathcal{H}_\varphi(\phi_k(\boldsymbol{y}))$. To make sure it obtains salient structure features and further filters out insignificant information, we apply a regression task $\mathcal{R}_\varphi$ on top of $h_k(\boldsymbol{y})$, and constraints the output to be closer to its sharp target $\phi_k(\boldsymbol{x})$. In this way, the guidance $h_k(\boldsymbol{y})$ at scale k is enforced to maintain information that is relevant to a sharp image, and suppress other signals that are input-specific (e.g., trace of blurs).

Finally, we incorporate the multiscale guidance $\{h_k(\boldsymbol{y})\}$ to the original diffusion UNet by adding the extracted representation to the feature map at the respective scale on the diffusion encoder (Fig 2) as an extra bias. To compensate for the difference in depth, at each corresponding scale, we apply a convolutional layer that has the same number of features as in the diffusion encoder. Detailed diagram is provided in the appendix.

3.3. Training loss

Our model is trained end-to-end with both a multiscale regression loss for optimizing the guidance network, and a denoising loss in icDPM. The regression loss is the mean squared error at each scale k defined as:

$$\mathcal{L}^k_{\text{guidance}} = \mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})\sim p_{\text{train}}} \|\mathcal{R}_\varphi(\mathcal{H}_\varphi(\phi_k(\boldsymbol{y}))) - \phi_k(\boldsymbol{x})\|_2, \tag{2}$$

where $\mathcal{H}_\varphi$ is the guidance feature extractor, and $\mathcal{R}$ is instantiated as a single convolutional layer that projects the guidance feature to the final output towards the clean image (as depicted in Fig. 3). The total regression loss is the average over different scales $\mathcal{L}_{\text{guidance}} = \sum_k \mathcal{L}^k_{\text{guidance}}$. Note that we do not use any additional downsampling/upsampling operation in the guidance network, so the spatial dimension remains the same at each scale. We empirically observe that the best performance is obtained by integrating three different scales with $k = 1, 2, 3$ with details discussed in Sec. 4.6.

By aggregating the information from the input image $\boldsymbol{y}$, and the multi-scale guidance $\{h_k(\boldsymbol{y})\}$, our icDPM $\mathcal{G}$ is trained by minimizing the denoising loss,

$$\begin{aligned}\mathcal{L}_{\text{DPM}} = \mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})\sim p_{\text{train}}} \mathbb{E}_{t\sim \text{Unif}(0,1)} \mathbb{E}_{\boldsymbol{\epsilon}\sim \mathcal{N}(0,\boldsymbol{I})} \\ \|\mathcal{G}_\theta(\boldsymbol{x}_t, \boldsymbol{y}, \{\mathcal{H}_\varphi(\phi_k(\boldsymbol{y}))\}, \alpha_t) - \boldsymbol{\epsilon}\|_1.\end{aligned} \tag{3}$$

The denoising model parameterized by θ predicts the noise $\boldsymbol{\epsilon}$, given the noisy corruption $\boldsymbol{x}_t$, the blurry input $\boldsymbol{y}$, the noise scheduler α_t as well as the proposed multi-scale guidance $\{\mathcal{H}_\varphi(\phi_k(\boldsymbol{y}))\}$. The total training loss $\mathcal{L} = \mathcal{L}_{\text{guidance}} + \mathcal{L}_{DPM}$, which is used to optimize the guidance network $\mathcal{H}$, the regression layer $\mathcal{R}$, and icDPM $\mathcal{G}$ in an end-to-end manner. During the inference, the model starts with a Gaussian noise, and iteratively recovers the clean image, conditioned on both the blurry input and the multiscale guidance at each denoising step.

4. Experiments

4.1. Setup and metrics

As motivated above, we are particularly interested in the model generalization of DPMs to unseen blurry data. Therefore, we set up our experiments under the scenario that the model will be only trained with synthetic paired dataset, and will be evaluated on a few unseen testing sets where the images may present different content and distortions than the in-domain data. To benchmark, we use the widely adopted motion deblurring dataset GoPro [51] as our training data, and assume Realblur-J [60], REDS [50] and HIDE [69] are representatives of unseen test sets.

In **GoPro** [69], 3214 pairs of blurry/clean training examples are provided for training, and 1111 images are held-out for evaluation. **Realblur-J** [60] is a recent realistic dataset mainly consisting of low-light scenes with motion blur with 980 test images provided. We consider it to present the largest domain gap with GoPro. **REDS** [50] presents a complimentary video deblurring dataset with more realistic motion blur. We follow [50, 6] and extract 300 validation images for the motion deblur test. **HIDE** [69] is the most commonly adopted dataset to test the model generalization ability trained from GoPro with 2025 test images.

4.2. Implementation details

Our framework is implemented in TensorFlow 2.0 and trained on 32 TPU v3 cores. We warm start the training with only regression loss, and linearly increase the weight of the denoising loss to 1 within the first 60k iterations. Adam optimizer [34] is used during the training ($\beta_1 = 0.5, \beta_2 = 0.999$), with batch size 256 on 128×128 random crops. We use linear increasing learning rate within the first 20k iterations, then with a constant learning rate 1×10^{-4}. We use a fully-convolutional UNet architecture [83] for icDPM to ensure the model can be used at arbitrary image resolutions. During the inference, we follow [83] and perform a sequence of sampling under different parameters. More details are included in the appendix.

4.3. Effectiveness of the guidance

We first validate the effectiveness of the proposed guidance module by qualitatively comparing with our baseline setup, which is a standard image-conditioned DPM (abbrev as 'icDPM'), on top of which we will introduce the guidance module (abbrev as 'icDPM w/ Guide').

Table 1. Inception distances analysis (domain-shift) between two different domains (GoPro v.s. Realblur-J) at different scales within different image space. At each scale, the guidance network output consistently reduces the gap compared to the downsampled input images, as expected. The distance between grayscale inputs (at original spatial resolution) are also given as a reference. KID values are scaled by a factor of 100 for readability.

Space	FID ↓	KID ↓
Input	61.115	3.07
Input ×2 downsampled	58.266	3.02
Guidance ×2 output	49.437	3.00
Input ×4 downsampled	56.313	3.60
Guidance ×4 output	47.984	3.46
Input ×8 downsampled	49.684	4.91
Guidance ×8 output	44.649	4.70

Guidance and domain gap. As the guidance is designed to improve the robustness towards domain shift, we perform an analysis on the Inception distances from different intermediate 'image space' to verify whether the guidance module is progressively reducing the gap between inputs from different sources (i.e. different blurry images in our scenario). In Table 1, we start by calculating the per-scale Inception distance between GoPro (in-domain) and Realblur-J (out-of-domain) images. At each scale of $\times 2$, $\times 4$ and $\times 8$ downsampled space, we observe a consistent reduction of FID and KID on the guidance network outputs, compared with the downsampled grayscale inputs. This demonstrates that the introduction of the learned guidance may provide more domain-agnostic information and benefit generalization of the model on unseen domains. In Fig. 4, we display the multiscale regression outputs at different scales on an out-of-domain input. The results align with our expectations, as the grayscale prediction at each scale progressively approached a clean image. Further, we noticed pronounced sampling artifacts from icDPM in this example, which are effectively eliminated with the proposed guidance.

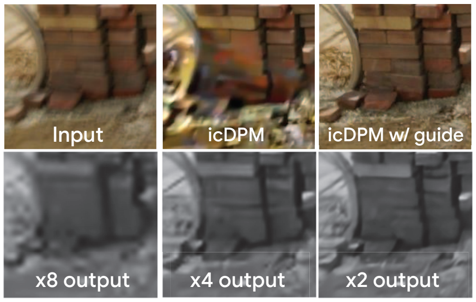

Figure 4. Top row: We compare the visual deblurring results on an out-of-domain image between a standard icDPM and icDPM with the proposed guidance module. While icDPM is prone to producing artifacts, our method that incorporates the guidance is more robust. Bottom row: We also visualize our multiscale regression outputs from scales of $\times 8$, $\times 4$, $\times 2$, indicating the prediction at spatial resolutions of $1/8, 1/4, 1/2$ the input image.

Guidance and model capacity. As the guidance network introduces more parameters, we investigate if its performance improvement is solely due to larger models. We perform a joint analysis of varying model size with and without the guidance network, and results are presented in Table 2. We refer results on the GoPro test set as 'In-domain' and on the Realblur-J dataset as 'Out-of-domain', using a single GoPro-trained model. We keep the number of building blocks constant and modulate the network size by changing only the number of convolutional filters. '-S' and '-L' indicate a smaller and larger models, respectively. We start with the image-conditioned DPM (icDPM) without the pro-

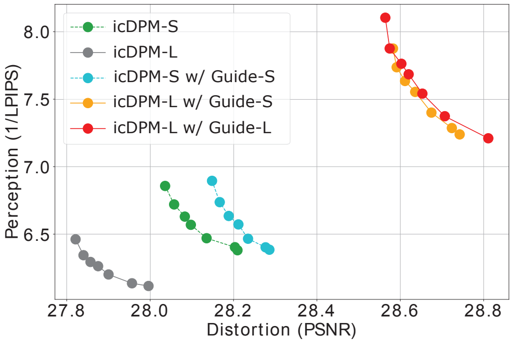

Figure 5. Perception-distortion plot as supplementary for Table 2, under varying sampling parameters. Under different network capacities ('-S' and '-L' refer to small and large respectively), the guidance mechanism allows for consistently better perceptual quality and lower distortions compared to icDPM. Plots for other datasets are in supplementary, where we observe similar trends.

Table 2. The effectiveness of the proposed guidance on top of image-conditioned DPM (icDPM), under different network size ('-S' and '-L' refer to small and large networks respectively). We show both In-Domain (train on GoPro, test on GoPro), and Out-of-Domain (train on GoPro, test on Realblur-J) results. Based on (a)-(b), we observe a larger icDPM boost in-domain performance, while not necessarily lead to better out-of-domain results. With guidance (c)-(d), we observe consistent improvements both in-domain and out-of-domain.

	Guidance network	Diffusion network	#Params	**In-Domain**		**Out-of-domain**	
				best LPIPS ↓	best PSNR ↑	best LPIPS ↓	best PSNR ↑
(a) icDPM-S	-	ch=32	6M	0.077	30.555	0.150	28.209
(b) icDPM-L	-	ch=64	27M	0.058	32.105	0.156	27.996
(c) icDPM-S w/ Guide-S	ch=32	ch=32	10M	0.068	31.298	0.145	28.286
(d) icDPM-L w/ Guide-S	ch=32	ch=64	30M	0.058	32.220	0.128	28.742
(e) icDPM-L w/ Guide-L	ch=64	ch=64	52M	0.057	32.254	0.123	28.711

Table 3. Average deblurring results across GoPro [51], HIDE [69] Realblur-J [60], REDS [50] dataset with GoPro [51]-only trained model, indicating the model robustness on various unseen data.

	Perceptual				**Distortion**	
	LPIPS ↓	NIQE ↓	FID ↓	KID ↓	PSNR ↑	SSIM ↑
DeblurGAN-v2 [37]	0.149	3.42	14.57	5.28	28.09	0.871
MPRNet [88]	0.140	3.70	20.22	8.49	29.78	0.897
UFormer [82]	0.133	3.65	18.99	8.13	30.06	0.903
Restormer [86]	0.139	3.69	19.90	8.36	30.00	0.895
Ours-SA	0.124	3.64	14.36	6.79	29.98	0.902
Ours	0.104	2.94	8.41	2.39	28.81	0.881

posed guidance network under different network sizes. In Table 2 row (a) and (b), we observe a significant improvement of the in-domain deblurring performance by increasing the UNet capacity, in terms of both perception and distortion qualities. However, the out-of-domain testing results become much worse with a larger network, suggesting potential overfitting during the training. Through visual inspection, we also found that the larger DPM is prone to artifacts when presented with unseen data, as shown in Fig. 1, 4 and 8. By introducing the guidance network, we observe both in-domain and out-of-domain performance gains. To isolate the effects of introducing guidance module, we also compare the performance between the icDPM and icDPM with guidance under similar amount of parameters, where we reduce the parameters of the model by using a smaller guidance module (from (e) to (d)). We observe comparable results of (d) icDPM-L w/ Guide-S with (e) icDPM-L w/ Guide-L in both in-domain and out-of-domain performance, and both of them outperformed icDPM-L.

We additionally present a distortion-perception plot in Fig. 5, with samples acquired from varying sampling parameters (i.e. number of steps and the standard deviation of noise). Similar to [83], we found a general trade-off between perceptual quality and distortion metrics. Also, we observed that all guided models consistently outperform the baseline DPMs under varying sampling parameters. We provide additional results on other datasets in appendix and observe similar effects and benefits of using the guidance module over the baseline icDPM.

4.4. Deblurring results

We compare our deblurring results with the state-of-the-art methods, loosely categorized into distortion-driven models [6, 88, 7], perception-driven GAN based methods [36, 37], as well as recent diffusion-based method [83]. In this work, we place particular emphasis on evaluating (1) the generalization ability on unseen data, and (2) the perceptual quality of the output, willing to compromise a slight drop on average distortion scores towards a better trade-off between distortion and perception [4]. For benchmarking, we mainly consider perceptual quality of the results quantified by standard metrics for generative models including LPIPS [94], NIQE [49], FID (Fréchet Inception Distance) [17], and KID (Kernel Inception Distance) [3]. We also present distortion metrics including PSNR and SSIM for completeness. However, we note that they are less correlated with human perception [53] and maximizing PSNR/SSIM results in a compromise of visual perception [4]. Our method is based on generative models and performs stochastic posterior sampling. The reference image provided in the training dataset is only one of the possible restoration result among other possibilities (due to the ill-posed nature of inverse problems). Thus, similar to [4, 55], our results compromise a certain amount of pixelwise average distortion while still being faithful to the target. We highlight best and second-best values for each metric. KID values are scaled by a factor of 1000 for readability.

We first present the in-domain GoPro performance in Table 4. Our model achieved state-of-the-art perceptual metrics across the board, while maintaining competitive distortion metrics by taking average of multiple samples ('Ours-SA'). Moreover, we are interested in the domain generalization and out-of-domain results on Realblur-J (Table 5), REDS (Table 7) and HIDE (Table 6). We achieved a significantly better perceptual quality on the unseen Realblur-J and REDS, and competitive results on HIDE. Further, we analyze the robustness the best-performing single-dataset trained models by comparing their average performance across all four test sets summarized in Table 3. Our method significantly improves the perceptual scores, while main-

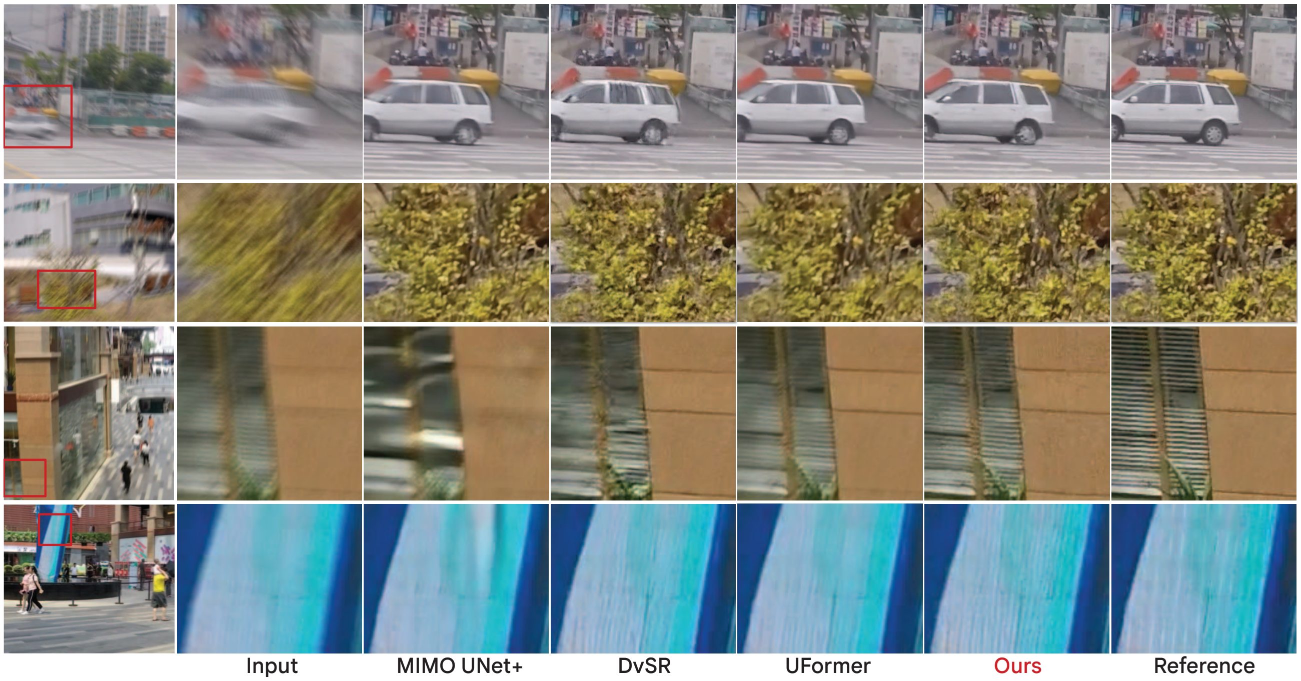

Figure 6. Deblurring results on GoPro [51] (top two rows) and HIDE [69] (bottom two rows) test set from MIMO UNet+ [7], DvSR [83], UFormer [82], and Ours. All models are trained only on GoPro [51] training set. Our method generates perceptually much sharper images, and reduce artifacts when applied to unseen images (HIDE). Enlarged results are provided in the appendix.

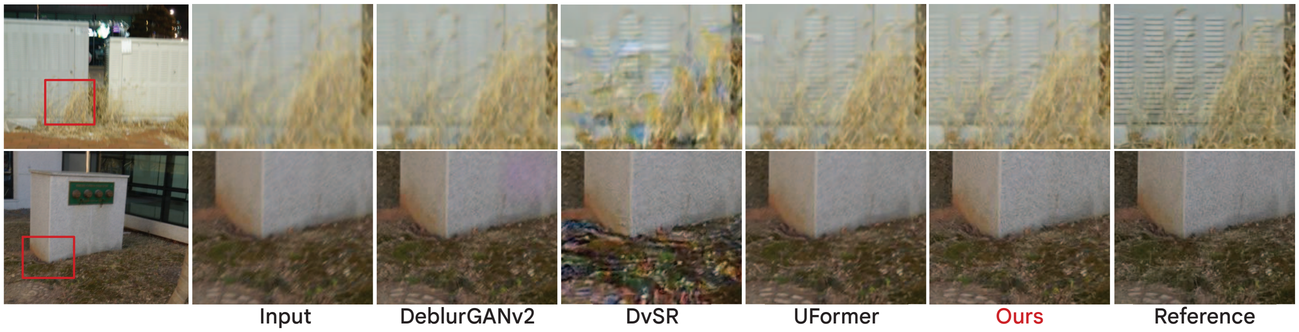

Figure 7. Test examples on Realblur-J [60], with all models trained only on GoPro dataset [51]. Best viewed electronically. We empirically found DeblurGANv2 [37] and DvSR [83] are prone to artifacts, and regression-based UFormer [82] produces over-smoothing images. Our methods alleviate artifacts, and produce high-fidelity deblurring on unseen data, even when the domain gap is large.

taining highly competitive distortion scores with $<$ 0.08 dB difference from the best PSNR, and $<$ 0.001 difference from the best SSIM with averaging samples ('Ours-SA').

Visual deblurring examples are provided in Fig. 6 on GoPro [51] and HIDE [69], Fig. 7 on RealblurJ [60] and Fig. 8 on REDS [50], respectively. On the GoPro (in-domain) test example, we find that all methods are able to produce reasonable artifact-free deblurring, and our method generates sharper and more visual realistic results. On the three out-of-domain datasets, performance degradation starts to occur from baseline methods. For instance, GAN-based model [37] and previous diffusion based model [83] tend to produce artifacts on out-of-domain data, and state-of-the-art regression based model [82] produces over-smoothed results. Our formulation performs more consistently better across different datasets, significantly reducing artifacts on unseen data with high perceptual realism. More enlarged visual examples are in the appendix.

4.5. Perceptual Study

We further conducted a user study with human subjects to verify the perceptual quality of the deblurring performance on unseen data, with all models trained on GoPro and tested on Relblur-J. We asked Amazon Mechanical Turk raters to select the best quality image from a given pair. We used 30 unique pairs of size 512×512 and averaged the

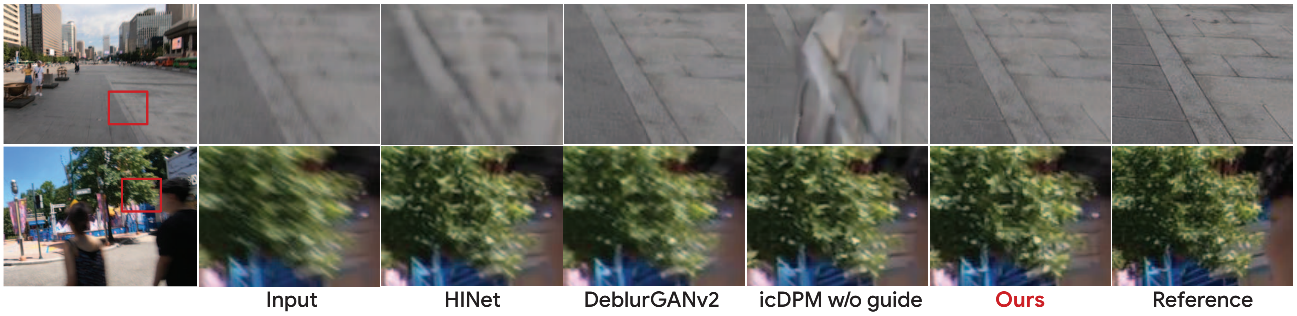

Figure 8. **REDS** [50] deblurring examples from HINet [6], DeblurGAN-v2 [37], icDPM without guidance and Ours. When trained only on GoPro [51], our method better removes the blur trace from the input image, while eliminating artifacts when applied on unseen data.

Table 4. Image deblurring results on GoPro [51] dataset.

	Perceptual				**Distortion**	
	LPIPS ↓	NIQE ↓	FID ↓	KID ↓	PSNR ↑	SSIM ↑
HINet [6]	0.088	4.01	17.91	8.15	32.77	0.960
MPRNet [88]	0.089	4.09	20.18	9.10	32.66	0.959
MIMO-UNet+ [7]	0.091	4.03	18.05	8.17	32.44	0.957
SAPHNet [77]	0.101	3.99	19.06	8.48	31.89	0.953
SimpleNet [43]	0.108	-	-	-	31.52	0.950
DeblurGANv2 [37]	0.117	3.68	13.40	4.41	29.08	0.918
DvSR [83]	0.059	3.39	4.04	0.98	31.66	0.948
DvSR-SA [83]	0.078	4.07	17.46	8.03	33.23	0.963
UFormer [82]	0.087	4.08	19.66	9.09	32.97	0.967
Restormer [86]	0.084	4.12	19.33	8.75	32.92	0.961
Ours-SA	0.078	4.10	8.69	7.06	33.20	0.963
Ours	0.057	3.27	3.50	0.77	31.19	0.943

Table 5. Results on Realblur-J [60] with GoPro trained models.

	Perceptual				**Distortion**	
	LPIPS ↓	NIQE ↓	FID ↓	KID ↓	PSNR ↑	SSIM ↑
UNet [63]	0.175	3.911	22.24	8.07	28.06	0.857
DeblurGAN [36]	-	-	-	-	27.97	0.834
DeblurGAN-v2 [37]	0.139	3.870	14.40	4.64	28.70	0.866
MPRNet [88]	0.153	3.967	20.25	7.57	28.70	0.873
DvSR [83]	0.153	3.277	18.73	6.00	28.02	0.851
DvSR-SA [83]	0.156	3.783	20.09	7.43	28.46	0.863
Restormer [86]	0.149	3.916	19.55	7.12	28.96	0.879
UFormer-B [82]	0.140	3.857	18.56	7.02	29.06	0.884
Ours-SA	0.139	3.809	16.84	6.25	28.81	0.872
Ours	0.123	2.976	12.95	3.58	28.56	0.862

Table 6. Results on HIDE [69] with GoPro [51] trained models.

	Perceptual				**Distortion**	
	LPIPS ↓	NIQE ↓	FID ↓	KID ↓	PSNR ↑	SSIM ↑
HINet [6]	0.120	3.20	15.17	7.33	30.33	0.932
MIMO-UNet+ [7]	0.124	3.24	16.01	7.91	29.99	0.930
MPRNet [88]	0.114	3.46	16.58	8.35	30.96	0.940
SAPHNet [77]	0.128	3.21	16.78	8.39	29.99	0.930
DeblurGAN-v2 [37]	0.159	2.96	15.51	6.96	27.51	0.885
DvSR-SA [83]	0.105	3.29	15.34	8.00	30.94	0.940
DvSR [83]	0.089	2.69	5.43	1.61	29.77	0.922
UFormer [82]	0.113	3.40	16.27	8.51	30.89	0.920
Restormer [86]	0.108	3.41	15.84	8.28	31.22	0.923
Ours-SA	0.104	3.40	14.62	7.62	30.96	0.938
Ours	0.088	2.91	5.28	1.68	29.14	0.910

Table 7. Results on REDS [50] with GoPro-trained models.

	Perceptual				**Distortion**	
	LPIPS ↓	NIQE ↓	FID ↓	KID ↓	PSNR ↑	SSIM ↑
HINet [6]	0.195	3.223	21.48	7.91	26.72	0.818
DeblurGAN-v2 [37]	0.181	3.172	14.98	5.12	27.08	0.814
MPRNet [88]	0.204	3.282	23.90	8.94	26.80	0.814
Restormer [86]	0.213	3.326	24.86	9.30	26.91	0.818
UFormer-B [82]	0.192	3.272	21.48	7.91	27.31	0.842
Ours-SA	0.178	3.248	17.27	6.21	26.95	0.834
Ours	0.147	2.610	11.91	3.51	26.36	0.810

Table 8. Perceptual study with human subjects on Realblur-J [60] dataset using GoPro [51] trained models. Each value represents the fraction of times that raters preferred the row over the column.

	DGANv2 [37]	UFormer [82]	DvSR [83]	icDPM	Ours-SA	Ours
DGANv2	-	0.28	0.43	0.56	0.26	0.17
UFormer	0.72	-	0.53	0.58	0.44	0.35
DvSR	0.57	0.47	-	0.61	0.32	0.29
icDPM	0.44	0.42	0.39	-	0.28	0.21
Ours-SA	0.74	0.56	0.68	0.72	-	0.38
Ours	0.83	0.65	0.71	0.79	0.62	-

750 ratings from 25 raters. In Table 8, each value represents the fraction of times that the raters preferred the row over the column. As can be seen, our method outperforms the existing solutions. Also, it is worth pointing out that there is a significant gap in the preference of our method with and without the guidance mechanism (denoted as icDPM).

4.6. Additional modeling choices

Guidance network. We carried out additional ablation studies for the modeling choices of the guidance network on regression target (RGB v.s. grayscale), the number of scales to adopt for the guidance (single v.s. multiscale), and the mechanism of incorporating the guidance (input-level vs latent space). For fair comparison, identical diffusion UNet is used under different configurations during training, and the same sampling parameters are used during inference. Table 9 (a) indicates our baseline icDPM without any guidance.We first compare the difference between incorporating the guidance at input-level and at latent space (Table 9 (b) and (c)). In (b), we upscale the regression output to the original input size, and concatenate the result to the diffusion UNet. In (c), we incorporate the feature maps before regression output into the UNet latent space via addition operation described above. The results indicate the benefit of latent-space guidance over input-level concatenation. Both (b)(c) improve on (a), showing the overall benefit of introducing the guidance. We also observe a moderate improvement by using multiscale guidance rather than single scale guidance in row (d) over (c). In row (e), we simplified the regression target from color space to grayscale space, which further improved the results.

Table 9. Effect of various settings on the domain invariant guidance. The scale column denotes the downsampling factors.

	Regression	Scale(s)	Guidance	LPIPS	PSNR
(a)	-	-	-	0.156	28.21
(b)	RGB	×8	input	0.145	28.34
(c)	RGB	×8	latent	0.143	28.45
(d)	RGB	×2, ×4, ×8	latent	0.141	28.45
(e)	Grayscale	×2, ×4, ×8	latent	0.137	28.63

Guidance operation. We performed ablation studies on the operations that injects the guidance into the icDPM backbone. We tested three different possible guidance operations including addition (our final choice), concatenation, and adaptive group normalization. Table 10 shows the deblurring PSNR on GoPro, where addition achieves the best performance. Empirically, we observe that adaptive norm in our use case tends to introduce artifacts. Compared with concatenation, addition is more memory-efficient, and ensures that the guidance cannot be simply neglected.

Table 10. Ablation on guidance operations.

Operation	PSNR
Addition	31.139
Concatenate	30.248
Adaptive norm	29.676

Training objectives for guidance feature. As we aim to enhance the icDPM backbone with a more robust conditioning mechanism through the integration of multiscale structure guidance, we introduce a regression loss to jointly train the guidance module and the icDPM. We assume that this loss is essential to ensure that the derived guidance representation retains prominent structural features while filtering out extraneous information. To verify our assumption, we trained our model without the regression loss, and we observed that the PSNR on GoPro test set degraded from 31.19 to 30.56. This indicates that the regression loss is crucial to effectively constraint learning of the guidance, and removing the loss potentially makes the model equivalent to just a larger UNet architecture.

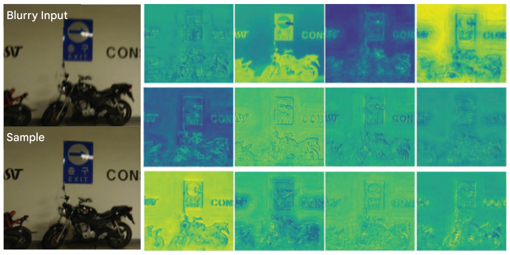

Figure 9. An out-of-domain deblurring result on a test image from Realblur-J [60] (left), alongside 12 (out of 64) selected channel-wise guidance feature maps at the scale of $k = 1$ (right).

Qualitatively, we examine the learned channel-wise guidance feature maps depicted in Fig. 9. As expected, these feature maps are highly related to edges and overall structures, which provide as auxiliary information to the icDPM about the coarse structures of its sharp reconstruction, and eventually benefit the robustness of the model.

5. Discussion

We present a learned multiscale structure guidance mechanism for icDPM that acts as an implicit bias which enhances its deblurring robustness. We acknowledge that limitations exist and require further investigation.

Although our focus is on improving the model's ability to generalize to unseen data without access to large-scale realistic training data, we recognize that the quality and realism of the training dataset ultimately bounds the deblurring capability of the model. In our experiments, we are restricted to the GoPro training dataset for benchmarking, which does not adequately cover all real-world scenarios, such as saturated regions with poor light conditions, light streaks at night. We observe that almost all methods fail on deblurring such images and we include failure cases in appendix. We believe our method can further benefit from large-scale diverse sets of training data.

Further, while the scope of our work is specifically on improving the robustness of deblurring in the context of icDPM where the deblurring task is cast as a *conditional generation problem* instead of a regression problem, similar ideas could be explored in different contexts, e.g., making a regression model more robust. However, as the guidance module is also trained with regression objectives, attaching it to a regression model may potentially lead to having a single regression model with increased number of parameters. Such formulation along with other extensions (e.g. plugging the guidance into other state-of-the-art regression backbone like transformers) require further investigations.

Acknowledgments Co-author Guido Gerig acknowledges support from the New York Center for Advanced Technology in Telecommunications (CATT).

References

[1] Jérémy Anger, Mauricio Delbracio, and Gabriele Facciolo. Efficient blind deblurring under high noise levels. In *2019 11th International Symposium on Image and Signal Processing and Analysis (ISPA)*, pages 123–128. IEEE, 2019. 2

[2] Arpit Bansal, Eitan Borgnia, Hong-Min Chu, Jie S Li, Hamid Kazemi, Furong Huang, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Cold diffusion: Inverting arbitrary image transforms without noise. *arXiv preprint arXiv:2208.09392*, 2022. 2

[3] Mikołaj Bińkowski, Dougal J. Sutherland, Michael Arbel, and Arthur Gretton. Demystifying MMD GANs. In *International Conference on Learning Representations*, 2018. 6

[4] 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

[5] Joan Bruna, Pablo Sprechmann, and Yann LeCun. Super-resolution with deep convolutional sufficient statistics. In *International Conference on Learning Representations*, 2016. 1, 2

[6] 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. 1, 2, 4, 6, 8

[7] 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, 6, 7, 8

[8] Hyungjin Chung, Byeongsu Sim, Dohoon Ryu, and Jong Chul Ye. Improving diffusion models for inverse problems using manifold constraints. *arXiv preprint arXiv:2206.00941*, 2022. 2

[9] Florinel-Alin Croitoru, Vlad Hondru, Radu Tudor Ionescu, and Mubarak Shah. Diffusion Models in Vision: A Survey. *arXiv preprint arXiv:2209.04747*, 2022. 2

[10] Giannis Daras, Yuval Dagan, Alexandros G Dimakis, and Constantinos Daskalakis. Score-guided intermediate layer optimization: Fast langevin mixing for inverse problem. *arXiv preprint arXiv:2206.09104*, 2022. 2

[11] Giannis Daras, Mauricio Delbracio, Hossein Talebi, Alexandros G Dimakis, and Peyman Milanfar. Soft diffusion: Score matching for general corruptions. *arXiv preprint arXiv:2209.05442*, 2022. 2

[12] 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. 1, 2

[13] Mauricio Delbracio, Hossein Talebei, and Pevman Milanfar. Projected distribution loss for image enhancement. In *2021 IEEE International Conference on Computational Photography (ICCP)*, pages 1–12. IEEE, 2021. 1, 2

[14] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021. 2

[15] 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. 2

[16] 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. 1

[17] 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. 6

[18] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 1, 2, 3

[19] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021. 2

[20] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *arXiv:2204.03458*, 2022. 2

[21] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei Efros, and Trevor Darrell. Cycada: Cycle-consistent adversarial domain adaptation. In *International conference on machine learning*, pages 1989–1998. Pmlr, 2018. 3

[22] Emiel Hoogeboom and Tim Salimans. Blurring diffusion models. *arXiv preprint arXiv:2209.05557*, 2022. 2

[23] Jungeon Kim Junyong Lee Seungyong Lee Sunghyun Cho Jaesung Rim, Geonung Kim. Realistic blur synthesis for learning image deblurring. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022. 1, 2, 3

[24] Seo-Won Ji, Jeongmin Lee, Seung-Wook Kim, Jun-Pyo Hong, Seung-Jin Baek, Seung-Won Jung, and Sung-Jea Ko. Xydeblur: Divide and conquer for single image deblurring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17421–17430, 2022. 2

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

[26] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016. 1, 2

[27] Zahra Kadkhodaie and Eero Simoncelli. Stochastic solutions for linear inverse problems using the prior implicit in a denoiser. *Advances in Neural Information Processing Systems*, 34:13242–13254, 2021. 2

[28] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *arXiv preprint arXiv:2206.00364*, 2022. 2

[29] Bahjat Kawar, Michael Elad, Stefano Ermon, and Jiaming Song. Denoising diffusion restoration models. In *Advances in Neural Information Processing Systems*, 2022. 2

[30] Bahjat Kawar, Roy Ganz, and Michael Elad. Enhancing diffusion-based image synthesis with robust classifier guidance. *arXiv preprint arXiv:2208.08664*, 2022. 2

[31] Bahjat Kawar, Jiaming Song, Stefano Ermon, and Michael Elad. Jpeg artifact correction using denoising diffusion restoration models. *arXiv preprint arXiv:2209.11888*, 2022. 2

[32] Bahjat Kawar, Gregory Vaksman, and Michael Elad. Snips: Solving noisy inverse problems stochastically. *Advances in Neural Information Processing Systems*, 34:21757–21769, 2021. 2

[33] Diederik Kingma, Tim Salimans, Ben Poole, and Jonathan Ho. Variational diffusion models. *Advances in neural information processing systems*, 34:21696–21707, 2021. 2

[34] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[35] Dilip Krishnan, Terence Tay, and Rob Fergus. Blind deconvolution using a normalized sparsity measure. In *CVPR 2011*, pages 233–240. IEEE, 2011. 2

[36] Orest Kupyn, Volodymyr Budzan, Mykola Mykhailych, Dmytro Mishkin, and Jiri Matas. Deblurgan: Blind motion deblurring using conditional adversarial networks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8183–8192. IEEE, 2018. 1, 2, 6, 8

[37] 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, 6, 7, 8

[38] Rémi Laumont, Valentin De Bortoli, Andrés Almansa, Julie Delon, Alain Durmus, and Marcelo Pereyra. Bayesian imaging using plug & play priors: when langevin meets tweedie. *SIAM Journal on Imaging Sciences*, 15(2):701–737, 2022. 2

[39] 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. 1, 2

[40] Anat Levin, Yair Weiss, Fredo Durand, and William T Freeman. Efficient marginal likelihood optimization in blind deconvolution. In *CVPR 2011*, pages 2657–2664. IEEE, 2011. 2

[41] Dasong Li, Yi Zhang, Ka Chun Cheung, Xiaogang Wang, Hongwei Qin, and Hongsheng Li. Learning degradation representations for image deblurring. In *European Conference on Computer Vision*, pages 736–753. Springer, 2022. 1, 2

[42] Haoying Li, Yifan Yang, Meng Chang, Shiqi Chen, Huajun Feng, Zhihai Xu, Qi Li, and Yueting Chen. Srdiff: Single image super-resolution with diffusion probabilistic models. *Neurocomputing*, 479:47–59, 2022. 1, 2

[43] 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, 8

[44] Boyu Lu, Jun-Cheng Chen, and Rama Chellappa. Unsupervised domain-specific deblurring via disentangled representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10225–10234, 2019. 3

[45] 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

[46] 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

[47] Lars Mescheder, Andreas Geiger, and Sebastian Nowozin. Which training methods for gans do actually converge? In *International conference on machine learning*, pages 3481–3490. PMLR, 2018. 1

[48] Tomer Michaeli and Michal Irani. Blind deblurring using internal patch recurrence. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part III 13*, pages 783–798. Springer, 2014. 2

[49] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal processing letters*, 20(3):209–212, 2012. 6

[50] Seungjun Nah, Sungyong Baik, Seokil Hong, Gyeongsik Moon, Sanghyun Son, Radu Timofte, and Kyoung Mu Lee. Ntire 2019 challenge on video deblurring and super-resolution: Dataset and study. In *CVPR Workshops*, June 2019. 3, 4, 6, 7, 8

[51] 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. 1, 2, 4, 6, 7, 8

[52] Seungjun Nah, Sanghyun Son, Jaerin Lee, and Kyoung Mu Lee. Clean images are hard to reblur: Exploiting the ill-posed inverse task for dynamic scene deblurring. In *International Conference on Learning Representations*, 2022. 1, 2, 3

[53] Jim Nilsson and Tomas Akenine-Möller. Understanding ssim. *arXiv preprint arXiv:2006.13846*, 2020. 6

[54] Ben Niu, Weilei Wen, Wenqi Ren, Xiangde Zhang, Lianping Yang, Shuzhen Wang, Kaihao Zhang, Xiaochun Cao, and Haifeng Shen. Single image super-resolution via a holistic attention network. In *European conference on computer vision*, pages 191–207. Springer, 2020. 2

[55] 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 IEEE/CVF International Conference on Computer Vision*, pages 1805–1813, 2021. 1, 2, 6

[56] Jinshan Pan, Zhe Hu, Zhixun Su, and Ming-Hsuan Yang. Deblurring text images via l0-regularized intensity and gradient prior. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2901–2908, 2014. 2

[57] Jinshan Pan, Deqing Sun, Hanspeter Pfister, and Ming-Hsuan Yang. Blind image deblurring using dark channel prior. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1628–1636, 2016. 2

[58] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 2

[59] Mengwei Ren, Neel Dey, James Fishbaugh, and Guido Gerig. Segmentation-renormalized deep feature modulation for unpaired image harmonization. *IEEE Transactions on Medical Imaging*, 40(6):1519–1530, 2021. 3

[60] 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. 1, 3, 4, 6, 7, 8, 9

[61] 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*, pages 10684–10695, 2022. 2

[62] 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*, pages 10684–10695, 2022. 2

[63] 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

[64] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. 2022. 4

[65] Chitwan Saharia, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. In *ACM SIGGRAPH 2022 Conference Proceedings*, pages 1–10, 2022. 1, 2, 3

[66] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022. 2, 4

[67] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 1, 2, 3

[68] Yuanjie Shao, Lerenhan Li, Wenqi Ren, Changxin Gao, and Nong Sang. Domain adaptation for image dehazing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2808–2817, 2020. 3

[69] 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. 4, 6, 7, 8

[70] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pages 2256–2265. PMLR, 2015. 2

[71] Gowthami Somepalli, Vasu Singla, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Diffusion art or digital forgery? investigating data replication in diffusion models. *arXiv preprint arXiv:2212.03860*, 2022. 2

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

[73] Yang Song, Conor Durkan, Iain Murray, and Stefano Ermon. Maximum likelihood training of score-based diffusion models. *Advances in Neural Information Processing Systems*, 34:1415–1428, 2021. 2

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

[75] Yang Song and Stefano Ermon. Improved techniques for training score-based generative models. *Advances in neural information processing systems*, 33:12438–12448, 2020. 2

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

[77] 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. 8

[78] Fu-Jen Tsai, Yan-Tsung Peng, Yen-Yu Lin, Chung-Chi Tsai, and Chia-Wen Lin. Stripformer: Strip transformer for fast image deblurring. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022. 1, 2

[79] Zhengzhong Tu, Hossein Talebi, Han Zhang, Feng Yang, Peyman Milanfar, Alan Bovik, and Yinxiao Li. Maxim: Multi-axis mlp for image processing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5769–5780, 2022. 1, 2

[80] Wei Wang, Haochen Zhang, Zehuan Yuan, and Changhu Wang. Unsupervised real-world super-resolution: A domain adaptation perspective. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4298–4307, 2021. 3

[81] Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. Recovering realistic texture in image super-resolution by deep spatial feature transform. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 606–615, 2018. 1, 2

[82] Zhendong Wang, Xiaodong Cun, Jianmin Bao, Wengang Zhou, Jianzhuang Liu, and Houqiang Li. Uformer: A general u-shaped transformer for image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17683–17693, June 2022. 1, 2, 6, 7, 8

[83] 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. 1, 2, 3, 5, 6, 7, 8

[84] Valentin Wolf, Andreas Lugmayr, Martin Danelljan, Luc Van Gool, and Radu Timofte. Deflow: Learning complex image degradations from unpaired data with conditional flows. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 94–103, 2021. 3, 4

[85] 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

[86] 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*, 2022. 2, 6, 8

[87] 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 *European Conference on Computer Vision*, pages 492–511. Springer, 2020. 2

[88] 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*, 2021. 1, 2, 6, 8

[89] 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*, pages 5978–5986, 2019. 2

[90] Jiawei Zhang, Jinshan Pan, Jimmy Ren, Yibing Song, Linchao Bao, Rynson WH Lau, and Ming-Hsuan Yang. Dynamic scene deblurring using spatially variant recurrent neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2521–2529, 2018. 1, 2

[91] 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. 3

[92] Lvmin Zhang and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. *arXiv preprint arXiv:2302.05543*, 2023. 2

[93] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 2

[94] 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. 6

[95] Xuaner Zhang, Qifeng Chen, Ren Ng, and Vladlen Koltun. Zoom to learn, learn to zoom. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3762–3770, 2019. 2

[96] Shangchen Zhou, Chongyi Li, and Chen Change Loy. Lednet: Joint low-light enhancement and deblurring in the dark. *arXiv preprint arXiv:2202.03373*, 2022. 3

[97] Shangchen Zhou, Jiawei Zhang, Jinshan Pan, Haozhe Xie, Wangmeng Zuo, and Jimmy Ren. Spatio-temporal filter adaptive network for video deblurring. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2482–2491, 2019. 1, 2