Image Deblurring with Domain Generalizable Diffusion Models

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

Abstract

Diffusion Probabilistic Models (DPMs) have recently been employed for image deblurring. DPMs are trained via a stochastic denoising process that maps Gaussian noise to the high-quality image, conditioned on the concatenated blurry input. Despite their high-quality generated samples, image-conditioned Diffusion Probabilistic Models (icDPM) rely on synthetic pairwise training data (in-domain), with potentially unclear robustness towards real-world unseen images (out-of-domain). In this work, we investigate the generalization ability of icDPMs in deblurring, and propose a simple but effective guidance to significantly alleviate artifacts, and improve the out-of-distribution performance. Particularly, we propose to first extract a multiscale domain-generalizable representation from the input image that removes domain-specific information while preserving the underlying image structure. The representation is then added into the feature maps of the conditional diffusion model as an extra guidance that helps improving the generalization. To benchmark, we focus on out-of-distribution performance by applying a single-dataset trained model to three external and diverse test sets. The effectiveness of the proposed formulation is demonstrated by improvements over the standard icDPM, as well as state-of-the-art performance on perceptual quality and competitive distortion metrics compared to existing methods.

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 [4, 5, 35, 50, 55, 72, 73, 77, 82, 84, 90] optimize distortion metrics such as PSNR, they often produce over-smoothed outputs that lack visual fidelity [3, 11, 34]. Therefore, along the line of perceptual-driven methods [22, 37], GANs [14] are leveraged for improved visual fidelity [31, 32]. However, GAN training suffers from instability, mode-collapse and artifacts [46], which may hamper the plausibility of the generated images.

Recently, DPMs [16] have further improve the photo-realism in a variety of imaging inverse problems [36, 60, 62, 77], by conditioning the DPM on the degraded estimation, typically via input-level concatenation. Formulated as image-to-image task, image-conditioned DPM (icDPM) often rely on pairwise training, and similar to any regression based method, the generalization ability to unseen domain is still under-explored.

In image restoration, curating ample paired and realistic training data is notoriously challenging. Typically, such training dataset is artificially generated 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. In this paper, we follow the literature in domain adaptation and transfer learning, and define the ‘domain gap’ as a distribution shift between the synthetic blurry inputs (in-domain) whose groundtruth labels are known, and the unseen/real blurry inputs (out-of-domain) whose groundtruth counterparts are unknown. Therefore, the synthetically trained models may
suffer from in-domain overfitting and may fail to generalize well to unseen data. As shown in Fig. 1, we compare results from recent state-of-the-art regression based [76, 82], GAN-based [32], and icDPM (our baseline) models. Similar to prior works, we observe a degraded performance of icDPM when applied to unseen data, and artifacts (also see example in Fig. 4). We speculate this may be related to the image-conditioning of current DPM [60, 62, 77], where image restoration is considered a form of image-to-image translation, and the degraded estimation is directly concatenated at the input-level of UNet, which implicitly results in domain-sensitivity. As pointed out by [57], conditioning DPM on blurred or corrupted images is still an under-explored research area. We claim that in image restoration, effective image conditioning is a key ingredient in improving domain generalization. The diffusion model should be conditioned on a signal that retains all the information relevant to generating the high-quality reconstruction, but it also requires the appropriate distilled information which ideally discards degradation and corruption from the source image.

In this work, we focus on improving domain generalization of conditional DPMs via more effective image conditioning, and make the following contributions: (1) we investigate and analyze the domain generalization of conditional diffusion models in motion deblurring task, and empirically find a relationship between generalization and image-conditioning; (2) we propose an intuitive but effective representation learning module that projects the input image to a multiscale domain-generalizable latent space, and incorporates such representation as auxiliary guidance for the diffusion network; (3) based on extensive ablation and benchmarking, we improve on the standard image-conditioned Diffusion Probabilistic Model (icDPM), and achieve state-of-the-art perceptual quality and competitive distortion metrics.

2. Related Works

Image 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 [4, 5, 21, 35, 50, 55, 72, 73, 75, 77, 82, 84, 90]. Pixel-wise losses, such as $L_1$ and $L_2$, are known to result in over-smoothed images [3, 11, 34] given their ‘regression to the mean’ nature. To this end, perception-driven losses including perceptual [11, 22, 43, 44, 86, 88] and adversarial losses [31, 32] are added on top of the pixel-wise constraints, to improve the visual fidelity of the deblurred image. Tangentially, recent works seek to improve the architectural design by exploring attention mechanisms [52, 72, 73, 76, 80, 81], multi-scale paradigms [5, 49] and multi-stage frameworks [4, 82, 83].

Diffusion Probabilistic Models (DPM) [12, 16, 65, 66]. Score-based models [68–70] and their recent exploratory generalizations [1, 10, 20] have achieved remarkable results in a varied range of applications [8], from image and video synthesis [17, 18, 26, 53, 58, 61], to solving general imaging inverse problems [6, 9, 23, 25, 28, 33]. DPMs are characterized for having a very stable training [12, 16, 24], diverse mode coverage [29, 67], and high synthesis fidelity [12, 53, 61]. 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 [12], text [53, 61], source image).

Image-conditioned DPM (icDPM) have been successfully re-purposed for image restoration tasks such as super-resolution [36, 62], deblurring [77], JPEG restoration [27, 60]. This is achieved by concatenating the corrupted observation to the noisy input. Such adaptations do not require task-specific losses or architectural designs, and have been adopted due to high sample perceptual quality. 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. Extensive efforts have been devoted to improving such generalization on unseen data via either curating more representative training data, or improving the domain generalization power of the model. To tackle the data limitations, previous works focus on acquiring or combining more representative training data [48, 55, 56, 89]. Another approach is to generate realistic degraded images using generative approaches [78, 85]. Under the umbrella of transfer learning, previous works consider explicit reduction of domain gap via unpaired image translation [19, 54] and domain adaptation [40, 63, 74]. Typically, these methods involve an adversarial formulation and require a joint training among synthetic and real domains, that may further complicate the training.

Unknown generalization of icDPMs. Since the existing image restoration icDPMs [36, 60, 62, 77] rely on similar synthetic training paradigm as that of regression models, we speculate that their generalization ability requires further investigation. In this work, we show that conditioning the diffusion model with domain-generalizable representations can significantly improve the quality of the restored images. We specifically focus on the deblurring problem, and develop a DPM-based restoration model that greatly benefits from an elegant conditioning method. The proposed framework improves upon the cross-domain performance of the existing
3. Method

3.1. Overview

We assume access to a paired dataset with samples \((x, y) \sim p_{\text{train}}(x, y)\), where \(x\) represents the high-quality sharp image, and \(y\) is the respective low-quality blurry observation. Such paired dataset is typically generated by simulating degraded images from the high-quality ones adopting a specific degradation model. The goal for image restoration is to reconstruct a clean, sharp image \(x\) from the low-quality observation \(y \sim p_{\text{real}}(y)\), coming from the distribution of real degraded images \(p_{\text{real}}(\cdot)\). In general, the distribution of the training dataset \(p_{\text{train}}\) differs from that of real degraded images \(p_{\text{real}}\).

DPMs We consider a general-purpose DPM for our formulation given its superior performance in high-quality image restoration [62, 77]. We use a fully-convolutional UNet architecture [77] to ensure the model can be used at arbitrary image resolutions. 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 \(p(x)\) 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. In the forward process depicted in Fig. 2, a noisy version \(x_t\) of the target image \(x\) is generated at a diffusion step \(t\) by \(x_t = \sqrt{1-t}x + (1-\sqrt{1-t})x, \epsilon \sim N(0, I_d)\), whereas in the reverse process, an image-to-image network (i.e., UNet) \(G_{a}(x_t, t)\) parameterized by \(\theta\) learns to estimate the clean image from the partially noisy input \(x_t\). In practice, a reparameterization of the model to predict the noise instead of the clean image leads to better sample quality [16].

Image-conditioned DPMs further inject an input image \(y\) so as to generate high-quality samples that are paired with the low-quality observation. Formally, this can be seen as generating samples from the conditional distribution of \(p(x|y)\) (posterior). This implies a conditional DPM \(G_{a}([x_t, y], t)\) where the image conditioning is typically implemented via concatenation of \(y\) and \(x_t\) at input-level [60, 62, 77]. Empirically, we found that such formulation is sensitive to domain-shift in the input images, and it leads to inferior performance on out-of-distribution image restoration (‘DPM’ in Fig. 1). Moreover, it may even produce visual artifacts (‘DvSR’ in Fig. 6). Therefore, we speculate that a more effective image-conditioning mechanism is required for boosting the robustness of image-conditioned DPM in the context of deblurring on unseen data. Intuitively, the DPM should be aware of certain domain-invariant priors such that it becomes less sensitive to the distribution shifts of inputs. We further assume the characteristics of such priors to be as domain-invariant as possible, meaning that they should discard information that reveals the input domain, while preserving the underlying image content. To this end, we propose a conditioning mechanism that contributes to providing the domain-generalizable guidance illustrated in Fig. 2, in order to reduce the sensitivity of the image-conditioned DPMs to the input domain.

3.2. Domain-generalizable DPM guidance

Fig. 3 shows the details of our proposed guidance module. Formally, we posit that the DPM equipped with such guidance module better learns to sample from the target conditional distribution to \(p_{\text{real}}(x|y)\).

Let us denote the guidance by \(h(y)\). Our goal is to design \(h(\cdot)\) such that the distribution of \(h(y)\) does not change significantly when the input domain changes. This makes the DPM less sensitive to the input \(y\). In other words, this guidance \(h(y)\) should preserve underlying information about the image \(y\), such as semantics and structure, while filtering certain amount of domain-specific clues (e.g., image color,
blur types). We exploit several standard transformations in image processing (i.e., colorspace conversion, image down-sampling and multiscale regression) and construct \( h(\cdot) \) to seek a balance between removal and preservation of information.

Taking the blurry input image \( y \), a transformation function \( \phi \) is applied for removing certain amount of information. We define \( \phi_k(y) \) as first converting \( y \) to grayscale space \( \tilde{y} \), followed by a down-sampling by a factor of \( 2^k \), where \( k = 1, 2, 3 \). Motivated by \[78\], we also add a small amount of Gaussian noise to further mask domain-specific blur, and make the guidance more robust. Thus,

\[
\phi_k(y) = d_{ik}(\tilde{y}) + n, \quad n \sim \mathcal{N}(0, \sigma^2 I). \tag{1}
\]

To make sure the guidance also includes as much content information as possible, we use a trainable network \( \mathcal{H}_\phi \) that further projects \( \phi_k(y) \) into the representation/latent space, \( h_k(y) = \mathcal{H}_\phi(\phi_k(y)) \). Ideally, \( h_k(y) \) should contain image information that is not relevant to the degradation. This information could be used for predicting \( \phi_k(x) \), that is, the respective clean counterpart at the same scale. Thus, we constraint the guidance network with a regression objective through a final convolution layer \( \mathcal{R}_\phi(h_k(\phi_k(y))) \). After extracting the multi-scale representation \( \{h_k(y)\} \), we incorporate the guidance to the original diffusion UNet by adding the extracted representation to the feature map at the respective scale on the diffusion encoder (Fig 2). 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 domain-generalizable representation loss and the denoising loss in DPM.

The DPM guidance loss is the mean squared error at each scale \( k \), leading to a per-scale regression loss as:

\[
\mathcal{L}^{\text{guidance}}_k = \mathbb{E}_{(x,y) \sim p_{\text{train}} \mathcal{H}_\phi(\phi_k(y))} ||\mathcal{R}_\phi(h_k(\phi_k(y))) - \phi_k(x)||, \tag{2}
\]

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

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

\[
\mathcal{L}_{\text{DPM}} = \mathbb{E}_{(x,y) \sim p_{\text{train}}} \mathbb{E}_{\gamma \sim p_\gamma} \mathbb{E}_{\epsilon \sim \mathcal{N}(0,1)} \|\mathcal{G}_\theta(\sqrt{\gamma}x + \sqrt{1-\gamma}\epsilon, \{\mathcal{H}_\phi(\phi_k(y))\}, y, \gamma) - y\|_1. \tag{3}
\]

The distribution \( p_{\gamma} \) is the one used in \[77\]. This denoising model parameterized by \( \theta \), tries to predict the noise \( \epsilon \), given the noisy corruption \( \tilde{x} \), the blurry input \( y \), as long as the proposed multi-scale guidance \( \{\mathcal{H}_\phi(\phi_k(y))\} \).

The overall training loss is \( \mathcal{L} = \mathcal{L}_{\text{guidance}} + \mathcal{L}_{\text{DPM}} \). Note that the guidance network is actually trained using information from the diffusion task, but also from the regression loss.

### 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 out-of-domain testing set where the images may present different content and distortions than the in-domain data. In order to compare with existing literature, we use the widely adopted motion deblurring dataset GoPro \[49\] as our synthetic training data, and assume Realblur-J \[56\], REDS \[48\] and HIDE \[64\] are representative of real unseen test set.

In GoPro \[64\], 3214 pairs of blurry/clean training examples are provided in the dataset, and 1111 images are held-out for evaluation. Realblur-J \[56\] 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. Therefore, a majority of visualizations in our work is focused on this dataset. We present visual results from \[48, 64\] in the appendix. REDS \[48\] presents a complimentary video deblurring dataset with more realistic motion blur. We follow \[4, 48\] and extract 300 validation images for the motion deblur test. HIDE \[64\] 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 \[30\] 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 \(1 \times 10^{-3}\). Detailed network architecture are included in the appendix. During
Guidance and domain gap. As the guidance module is an additional module (abbrev as 'icDPM w/ Guide'), on top of which we will introduce the guidance network outputs, which is a standard image-conditioned DPM (abbrev 'icDPM'), we perform an analysis on the Inception distance between GoPro (in-domain) and Realblur-J (out-of-domain) images. At each scale of ×2, ×4 and ×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. Empirically, we found the results match our expectation, where at each scale, the grayscale prediction is closer to a clean image. On this example, we observe strong sampling artifacts from icDPM, which are eliminated with the proposed guidance.

Guidance and model capacity. As introducing the guidance network also increases the number of parameters, it is likely that the improvements in performance are solely due to the benefit of larger models. We therefore present a joint analysis of varying model size with or without the auxiliary guidance network, with results reported in Table 2. We use 'In-domain' to present the results of GoPro deblurring with a GoPro trained model and 'Out-of-domain' as GoPro trained model on Realblur-J. We use the same number of building blocks and modulate the network size by only changing the number of convolutional filters. ‘-S’ and ‘-L’ indicate a smaller and larger models, respectively.

As our baseline, we start with the image-conditioned DPM (icDPM) without the proposed guidance network under different network sizes. In rows (a) and (b) of Table 2, we observe a significant improvement of the in-domain deblurring performance by increasing the UNet capacity, in both perception and distortion qualities. However, the out-of-domain testing results become much worse. Although a larger network in principle provides a higher modeling capacity, we speculate a potential overfitting happened during the training. Through visual inspection, we also found that the larger DPM is prone to artifacts as shown in Fig. 4. By introducing the guidance network, we observe both in-domain and out-of-domain performance gains. 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 [77], 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.

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’).

Guidance and domain gap. As the guidance module is motivated by reducing domain gap and improving domain invariance, 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 ×2, ×4 and ×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. Empirically, we found the results match our expectation, where at each scale, the grayscale prediction is closer to a clean image. On this example, we observe strong sampling artifacts from icDPM, which are eliminated with the proposed guidance.

Table 1. Inception distances analysis (domain-shift) between two different domains (GoPro v.s. Realblur-J) at different scales without 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 predictions 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  |

4.4. Deblurring results

We compare our deblurring results with the state-of-the-art methods, loosely categorized into distortion-driven mod-
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)-(f), we observe consistent improvements both in-domain and out-of-domain.

| Guidance network | In-Domain | Out-of-domain |
|------------------|------------|---------------|
| (a) icDPM-S      | ch=32     | ch=32        |
| (b) icDPM-L      | ch=64     | ch=64        |
| (c) icDPM-S w/ Guide-S | ch=32 | ch=32 | 10M | 0.068 | 31.298 | 0.145 | 28.286 |
| (d) icDPM-L w/ Guide-L | ch=64 | ch=32 | 30M | 0.058 | 32.220 | 0.128 | 28.742 |
| (e) icDPM-L w/ Guide-S | ch=32 | ch=64 | 30M | 0.058 | 32.220 | 0.128 | 28.742 |
| (f) icDPM-L w/ Guide-L | ch=64 | ch=64 | 52M | 0.057 | 32.254 | 0.123 | 28.711 |

Figure 5. Perception-distortion plot as supplementary for Table 2, under varying sampling parameters. The guidance mechanism allows for consistent better perceptual quality and lower distortions compared to icDPM under different network capacity (‘-S’ and ‘-L’ refer to small and large respectively).

4.5. Perceptual Study

We conducted a perceptual study with human subjects to compare the proposed model with existing methods. The model is trained on GoPro, and tested on Realblur-J. We achieved a significantly better perceptual quality on the unseen Realblur-J and REDS, and competitive results on HIDE.

Fig. 6 shows representative deblurring results on Realblur-J test set, which we perceive to have the largest distribution gap with GoPro dataset. We notice that GAN-based model [32] and previous diffusion based model [77] tend to produce artifacts on out-of-domain data (which does not occur in-domain), and state-of-the-art regression based model [76] produces over-smoothed results. Our formulation largely reduces artifacts on unseen data, while maintaining image sharpness. More visual examples are provided in the appendix.
quality image from a given pair. Our results are shown in Table 7. Each value in the table represents the fraction of times that the raters preferred the row over the column. We used 30 image crops of size $512 \times 512$ and averaged the ratings from 25 raters. As can be seen, our method with guidance outperforms the existing solutions. We observed that raters may only prefer sample averaging (Ours-SA) in images with low amount of variations and textures. 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

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), as well as the mechanism of incorporating the guidance (input-level vs latent space). We use smaller models (row (c) in Table 2) as well as the same sampling parameters for faster prototyping and fair comparison. As the design of the guidance module is motivated by domain generalization, we anchor our comparison mainly on out-of-domain performance. Results are shown in Table 8. Row (a) indicates our baseline icDPM without any guidance, under the same sampling parameter. We first compare the difference between incorporating the guidance at input-level and at latent space (row (b) and (c) in Table 8). In (b), we upscale the regression output to the original input size, and concatenate the result to the diffusion UNet (leading to the number of input channels from 6
Table 4. Cross-domain performance on Realblur-J [56] datasets, with models trained only on GoPro.

| Method          | LPIPS | NIQE | FID  | KID  | PSNR | SSIM |
|-----------------|-------|------|------|------|------|------|
| Ours            | 0.175 | 3.911| 22.24| 8.07 | 28.06| 0.857|
| DeblurGAN [31]  | -     | -    | -    | -    | -    | -    |
| DeblurGAN-v2    | 0.139 | 3.870| 14.40| 4.64 | 28.70| 0.866|
| MPRNet [82]     | 0.153 | 3.967| 20.25| 7.57 | 28.70| 0.873|
| DvSR [77]       | 0.153 | 3.277| 18.73| 6.00 | 28.02| 0.851|
| DvSR-SA [77]    | 0.156 | 3.783| 20.09| 7.43 | 28.46| 0.863|
| Restormer [80]  | 0.149 | 3.916| 19.55| 7.12 | 28.96| 0.879|
| UFormer-B [76]  | 0.140 | 3.857| 18.56| 7.02 | 29.06| 0.884|
| Ours-SA         | 0.123 | 2.976| 12.95| 3.58 | 28.56| 0.862|

Table 5. Cross-domain performance on REDS [48] datasets, with models trained only on GoPro.

| Method          | LPIPS | NIQE | FID  | KID  | PSNR | SSIM |
|-----------------|-------|------|------|------|------|------|
| Ours            | 0.107 | 3.40  | 14.62| 7.62 | 29.62| 0.920|
| Restormer [80]  | 0.108 | 3.41  | 15.84| 8.28 | 30.89| 0.920|
| UFormer [76]    | 0.113 | 3.40  | 16.27| 8.51 | 30.89| 0.920|
| DvSR-SA [77]    | 0.105 | 3.29  | 15.34| 8.00 | 30.94| 0.938|
| Ours-SA         | 0.088 | 2.91  | 15.62| 7.62 | 30.96| 0.938|

Table 6. Image deblurring results on the HIDE [64] dataset.

| Method          | LPIPS | NIQE | FID  | KID  | PSNR | SSIM |
|-----------------|-------|------|------|------|------|------|
| Ours            | 0.120 | 3.46  | 14.38| 7.62 | 29.62| 0.920|
| Restormer [80]  | 0.124 | 3.24  | 16.01| 8.35 | 29.99| 0.930|
| UFormer [76]    | 0.120 | 3.32  | 15.68| 8.39 | 29.99| 0.930|
| DvSR-SA [77]    | 0.125 | 3.29  | 15.34| 8.00 | 30.94| 0.940|
| DvSR [77]       | 0.089 | 2.69  | 5.43 | 1.61 | 29.77| 0.922|
| UFormer [76]    | 0.113 | 3.40  | 16.27| 8.51 | 30.89| 0.920|
| Restormer [80]  | 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. Perceptual study with human subjects on Realblur-J [56] dataset using models trained on the GoPro dataset [49]. Each value represents the fraction of times that Amazon Mechanical Turk raters preferred the row over the column. Results are averages over 750 ratings from 25 raters and 30 unique image pairs.

| Model          | LPIPS | NIQE | FID  | KID  | PSNR | SSIM |
|----------------|-------|------|------|------|------|------|
| DGANv2         | 0.140 | 3.87  | 16.84| 6.25 | 28.81| 0.872|
| UFormer        | 0.140 | 3.85  | 18.56| 7.02 | 28.70| 0.866|
| DvSR           | 0.135 | 3.78  | 17.83| 6.00 | 28.02| 0.851|
| DvSR-SA        | 0.136 | 3.78  | 18.09| 7.43 | 28.46| 0.863|

Table 8. 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      | x 8      | 0.145| 28.34|
| (c)            | RGB      | x 8      | 0.143| 28.45|
| (d)            | RGB      | x 2, x 4, x 8 | 0.141| 28.45|
| (e)            | Grayscale| x 2, x 4, x 8 | 0.137| 28.63|

5. Discussion

We present a learned domain generalizable representation and conditioning mechanism for image-conditioned diffusion models that improves out-of-domain performance in image deblurring. It also opens up many follow-up questions which will be tackled in future works.

While we focus on improving the model generalization towards unseen data (assuming limited access to large-scale realistic training data), we found that the deblurring capability of the model is ultimately bounded by the quality and representativity of the training dataset. In particular, when there exists a strong gap between the training and testing domains, the generalization may be limited. In our experiments, we are restricted to the GoPro training dataset for benchmarking, which does not cover enough real scenes that may occur in Realblur-J. For instance, saturated regions with poor light conditions (e.g. light streaks at night) are not well represented in GoPro. We observe that almost all methods fail on deblurring such images and we include failure cases in appendix.

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 found that simplifying the regression target from color space to grayscale space further improves the results. We speculate that this is due to the fact that the color information is already contained in the input image, and it may not be necessary to additionally inject it in the guidance module.

To 9). 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) and (c) improve on (a), showing the overall potential benefit of introducing the guidance.

Table 6. Image deblurring results on the HIDE [64] dataset.
References

[1] 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

[2] Mikolaj B blowingowski, Dougal J. Sutherland, Michael Arbel, and Arthur Gretton. Denoising MMGANs. In International Conference on Learning Representations, 2018. 6

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

[4] 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, 18, 19, 20, 21

[5] 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, 4, 6, 8, 18, 22, 23, 24, 25, 26, 27, 28, 29

[6] 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

[7] Hyungjin Chung, Byeongsu Sim, and Jong Chul Ye. Come-closer-diffuse-faster: Accelerating conditional diffusion models for inverse problems through stochastic contraction. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 12413–12422, 2022. 17

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

[9] 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. 4

[10] 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

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

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

[13] Tim Dockhorn, Arash Vahdat, and Karsten Kreis. Genie: Higher-order denoising diffusion solvers. arXiv preprint arXiv:2210.05475, 2022. 17

[14] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherijl Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. Advances in neural information processing systems, 27, 2014. 1

[15] 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

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

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

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

[19] 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. 2

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

[21] 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

[22] 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

[23] 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

[24] 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

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

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

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

[28] Bahjat Kawar, Gregory Vaksman, and Michael Elad. Snips: Solving noisy inverse problems stochastically. Advances in Neural Information Processing Systems, 34:21757–21769, 2021. 2
Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 4

Diederik Kingma, Tim Salimans, Ben Poole, and Jonathan Gilmer. GANs trained by a two time-scale update rule converge to a local NCE equilibrium. *arXiv preprint arXiv:1701.04862*, 2017. 2

Orest Kupyn, Volodymyr Budzan, Mykola Mykhailych, Dmytro Mishkin, and Jiří 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

Orest Kupyn, Tetiana Martynyuk, 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, 17, 18, 19, 20, 21, 30, 31, 32, 33

Rémi Laumont, Valentín De Bortoli, André Almansa, Julie Delon, Alain Durmus, and Marcelo Pereyra. Bayesian imaging using plug & play: when langevin meets tweede. *SIAM Journal on Imaging Sciences*, 15(2):701–737, 2022. 2

Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tefani, 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

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

Haoying Li, Yifan Yang, Meng Chang, Shiqi Chen, Huajun Li, and Jun Zhu. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. *arXiv preprint arXiv:2206.00927*, 2022. 17

Hengyu Ma, Li Zhang, Xiatian Zhu, and Jianfeng Feng. Accelerating score-based generative models with preconditioned diffusion sampling. In *European Conference on Computer Vision*, pages 1–16. Springer, 2022. 17

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

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

Chenlin Meng, Ruiqi Gao, Diederik P. Kingma, Stefano Ermon, Jonathan Ho, and Tim Salimans. On distillation of guided diffusion models. *arXiv preprint arXiv:2210.03142*, 2022. 17

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

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

Seungjun Nah, Sungyong Baik, Seokil Hong, Gyeongsik Moon, Sanghyun Son, Radu Timofte, and Kyoung Mu Lee. Nitre 2019 challenge on video deblurring and super-resolution: Dataset and study. In *CVPR Workshops*, June 2019. 2, 4, 8, 13, 17, 18, 19, 20, 21

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, 13, 14, 26, 27, 28, 29

Seungjun Nah, Sanghyun Son, Jaerim Lee, and Kyoung Mu Lee. Clean images are hard to reblur: Exploiting the implicit inverse task for dynamic scene deblurring. In *International Conference on Learning Representations*, 2021. 1, 2

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

Ben Niu, Weilei Wen, Wenqi Ren, Xiangde Zhang, Liaping Yang, Shuzhen Wang, Kaihao Zhang, Xiaochnun 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

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

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. 2

Jaesung Rim, Geonung Kim, Jungeun Kim, Junyong Lee, Seungyong Lee, and Sunghyun Cho. Realistic blur synthesis for learning image deblurring. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022. 1, 2
[56] Jaesung Rim, Haeyun Lee, Jucheol Won, and Sunghyun Cho. Real-world blur dataset for learning and benchmarking de-blurring algorithms. In European Conference on Computer Vision, pages 184–201. Springer, 2020. 1, 2, 4, 7, 8, 13, 14, 30, 31, 32, 33, 34, 35, 36, 37

[57] 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

[58] 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

[59] 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. 8, 30, 31, 32, 33

[60] 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

[61] 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

[62] 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

[63] Yuanjie Shao, Lerenhan Li, Wenqi Ren, Changxin Gao, and Chitwan Saharia, Jonathan Ho, Tim Salimans, David Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 2808–2817, 2020. 2

[64] Ziyi Shen, Wenguan Wang, Jianbing Shen, Haibin Ling, Tingfu Xu, and Ling Shao. Human-aware motion deblurring. In IEEE International Conference on Computer Vision, 2019. 4, 8, 13, 22, 23, 24, 25

[65] 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

[66] Jianming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In International Conference on Learning Representations, 2021. 2, 17

[67] 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

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

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

[70] 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

[71] 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. 6, 8, 22, 23, 24, 25, 26, 27, 28, 29

[72] 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

[73] 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

[74] 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. 2, 15

[75] 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

[76] Zhengdong 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, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33

[77] 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, 4, 5, 6, 7, 8, 16, 17, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33

[78] 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. 2, 4

[79] Zhisheng Xiao, Karsten Kreis, and Arash Vahdat. Tackling the generative learning trilemma with denoising diffusion GANs. In International Conference on Learning Representations (ICLR), 2022. 17

[80] 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, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33

[81] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. Learning enriched features for real image restoration and enhancement. In European Conference on Computer Vision, pages 492–511. Springer, 2020. 2

[82] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. Multi-stage progressive image restoration. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021. 1, 2, 6, 8, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33

[83] 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

[84] 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

[85] Kaihao Zhang, Wenhan Luo, Yiran Zhong, Lin Ma, Bjorn Stenger, Wei Liu, and Hongdong Li. Deblurring by realistic blurring. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 2737–2746, 2020. 2

[86] Richard Zhang, Phillip Isola, Alexei 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

[87] Richard Zhang, Phillip Isola, Alexei 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

[88] 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

[89] Shangchen Zhou, Chongyi Li, and Chen Change Loy. Lednet: Joint low-light enhancement and deblurring in the dark. arXiv preprint arXiv:2202.03373, 2022. 2, 14

[90] 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
Appendix

A. Additional Results

A.1. Effectiveness of the guidance on GoPro, HIDE and REDS

We include additional perception-distortion plots for GoPro [49], HIDE [64] and REDS [48] datasets in Fig. 7, as supplementary for Section 4.3 of the main paper, to verify the effectiveness of the proposed guidance.

![Graphs showing perception-distortion plots for GoPro, HIDE, and REDS datasets.](image)

Figure 7. Additional perception-distortion plots as supplementary for Sec. 4.3 in the main paper. All models are trained only on GoPro [49]. The guidance mechanism allows for consistent better perceptual qualities and lower distortions compared to image-conditioned diffusion probablistic model (icDPM) under different network capacities (‘-S’ and ‘-L’ refer to small and large respectively), both in-domain (GoPro) and out-of-domain (HIDE, REDS). We also observe when the domain gap is mild (i.e. HIDE, REDS) compared to Realblur-J [56], using a smaller guidance network may lead to better perceptual-distortion balance than a large guidance network.

A.2. Additional visual results

Due to limited space, we only present visual examples from Realblur-J [56] in the main paper, which has the largest domain gap with the GoPro [49] training set. For completeness, we provide more qualitative results for all datasets below.

Realblur-J (out-of-domain) deblurring examples are shown in Fig. 23, 24, 25, 26.
REDs (out-of-domain) deblurring examples are shown in Fig. 11, 12, 13, 14.
HIDE (out-of-domain) deblurring examples are shown in Fig. 15, 16, 17, 18.
GoPro (in-domain) deblurring examples are shown in Fig. 19, 20, 21, 22.
A.3. Failure cases

As mentioned in the main paper, the domain generalization of the model is still extensively bounded by the quality of the training set. In our experiments, we only train with GoPro [49] for the sake of benchmarking. However, the data diversity and representativity from GoPro is limited, i.e., it only contains daytime scenes, acquired outdoor under sufficient lighting conditions. Moreover, the synthesis of blur in GoPro by simple averaging of consecutive frames is less realistic [89]. Lastly, the ground truth images in GoPro dataset are rather low-quality, which may further hurt the out-of-domain performance. Therefore, it is expected that it will be extremely hard for the model to perform decent deblurring on scenes significantly different from GoPro, such as low-light images with saturated regions, in Realblur-J [56].

We include a few failure cases on such scenes in Fig. 27, 28, 29 30, where all methods fail to remove blur from the night scenes, especially with night streaks. We believe that in practice, more realistic training datasets are needed to further increase the model generalization.

A.4. Guidance feature

We qualitatively analyze the channelwise guidance feature maps in Fig. 8. In general, these feature maps are qualitatively related to edges and overall semantics towards a clean image. We expect that using such features as additional guidance is beneficial for improved domain generalization, as verified by the qualitative and quantitative deblurring results in the main text.
B. Additional Ablation

Input concatenation During prototyping, we also explored the possibility of removing input-level concatenation, and only rely on the intermediate representations from regression as the condition of the diffusion model, similar as in [36] for super-resolution. Potentially, we expect such setting will further make the model domain-generalizable as it does not directly interact with images from different domains, although it may also risk losing detailed information from the input.

As proof of concept, we use the same multiscale regression networks, and compare the models with or without input concatenation. Further, since the diffusion model now only takes the intermediate representations as input, we reintroduce the RGB information by using our model variants (d) in Table 8. in the main paper (i.e., regression targets are downsampled RGB images instead of grayscale images). From Table 9, we observe that the input concatenation obtained a much better performance both in-domain and out-of-domain than without concatenation. Therefore, in our final model, we keep the input concatenation and only rely on the guidance features to provide additional information.

Table 9. Effects of input-level concatenation. From our model variant (d) in Table 8. of the main paper, we remove the input concatenation, loosely inspired by [36] (super-resolution). In the context of deblurring, we observe deteriorate results indicated in row ‘w/o input concatenation’, compared to the setting with additional input concatenation.

|                 | In-domain | Out-of-domain |
|-----------------|-----------|---------------|
|                 | PSNR ↑    | LPIPS ↓       | PSNR ↑    | LPIPS ↓   |
| w/o input concatenation | 25.20 0.230 | 28.29 0.177   | 28.45 0.141 |
| w/ concatenation   | 30.65 0.090 | 28.45 0.141   |

Further cross-domain alignment. We also explored the potential effects of finetuning the DPMs with adversarial formulation where we used additional discriminators on the guidance features between different datasets (e.g., GoPro and Realblur-J) so that the features extracted from different domains become indistinguishable, similar to the feature alignment strategy in [74]. However, we do not observe extra benefits, and find that such finetuning may even hurt the performance as shown in Fig. 9. We speculate that it could be a result of training instability of GANs, or perhaps the suboptimal formulation under the image-conditioned DPM framework. We will leave this for future investigation.

![Perception (1/LPIPS) vs Distortion (PSNR)](image)

Figure 9. A comparison between our models with or without further domain adaptation with Realblur-J, on a GoPro trained model. Surprisingly, further adversarial domain adaptation on the guidance features between GoPro and Realblur-J hurt the performance.
C. Additional implementation details

C.1. Architectures

The architectural details for the diffusion network and the guidance network are illustrated in Fig. 10.

Figure 10. The detailed architecture of the proposed method. **Left:** the image-conditioned diffusion network based on a fully-convolutional UNet similar to [77], where we replace the residual blocks from the UNet encoder with the proposed guided residual block. **Middle** column illustrates the difference between a standard residual block and the proposed guided block, where we additionally incorporate a guidance feature which is domain-generalizable. **Right:** The proposed guidance network for extracting the domain-generalizable guidance. At each scale, the blurry image is first converted to grayscale, downsampled, and lastly fed into the network to predict its clean counterpart. The output from the last residual block is leveraged as the guidance feature.
C.2. Inference

As we use continuous noise level sampling during training, it enables the use of different noise schedulers during the inference to potentially obtain samples with different distortion-perception trade-off. We therefore perform a grid search over a set of different diffusion steps $T$, as well as the upper bound of the noise variance $1 - \alpha_T$. For efficiency, we also exclude certain combinations that do not produce reasonable sampling (i.e., sampling results are pure noise or blank image), and the final combinations are indicated in Table 10.

Table 10. The sampling parameters for inference.

| Steps ($T$) | Maximum noise variance $1 - \alpha_T$ |
|------------|--------------------------------------|
|            | 0.01 0.02 0.05 0.1 0.2 0.5           |
| 20         | ✓                              |
| 30         | ✓                              |
| 50         | ✓ ✓                             |
| 100        | ✓ ✓ ✓                            |
| 200        | ✓ ✓ ✓ ✓                          |
| 500        | ✓ ✓ ✓ ✓                          |
| 1000       | ✓ ✓ ✓ ✓                          |

C.3. Computational cost

In Table 11, we report floating point operations per second (FLOPs) under different model configurations, calculated based on an input image of $720 \times 1280 \times 3$. For diffusion networks (c)-(d), the FLOPs are calculated based on a single diffusion step. While optimizing sampling speed is out-of-scope of this work, we believe recent advance in speeding up DPM sampling [7, 13, 38, 39, 41, 42, 45, 66, 79] could be further incorporated into our framework.

Table 11. FLOPs under different model configurations, calculated based on a full-size input image of $720 \times 1280 \times 3$. For diffusion networks (c)-(d), the FLOPs are calculated based on a single diffusion step.

| Guidance network | Diffusion network | # Params | FLOPs   |
|------------------|-------------------|---------|---------|
| (a) icDPM-S      | -                 | 6M      | 1200B   |
| (b) icDPM-L      | -                 | 27M     | 4800B   |
| (c) icDPM-S w/ Guide-S | ch=32     | 10M     | 2500B   |
| (d) icDPM-S w/ Guide-L | ch=32     | 30M     | 6400B   |
| (e) icDPM-L w/ Guide-S | ch=32     | 30M     | 6100B   |
| (f) icDPM-L w/ Guide-L | ch=64     | 52M     | 10000B  |

C.4. Benchmark results

We performed a consistent computation over all benchmarks for fair comparisons. To acquire the benchmark results, we use the author provided results whenever possible. On the cross-domain set up of Realblur-J with GoPro trained only models, we use author provided results of DvSR [77], UFormer [76], Restormer [80]. For DeblurGAN-v2 [32] and MPRNet [82], we use official code repository along with the provided GoPro checkpoints for inference. On REDS [48], all results are obtained by running their official models with the GoPro checkpoints.
Figure 11. REDS [48] deblurring examples from MPRNet [82], HINet [4], DeblurGAN-v2 [32], Restormer [80], UFormer [76], icDPM without guidance and Ours (icDPM with guidance).
Figure 12. REDS [48] deblurring examples from MPRNet [82], HINet [4], DeblurGAN-v2 [32], Restormer [80], UFormer [76], icDPM without guidance and Ours (icDPM with guidance).
Figure 13. REDS [48] deblurring examples from MPRNet [82], HINet [4], DeblurGAN-v2 [32], Restormer [80], UFormer [76], icDPM without guidance and Ours (icDPM with guidance).
Figure 14. REDS [48] deblurring examples from MPRNet [82], HINet [4], DeblurGAN-v2 [32], Restormer [80], UFormer [76], icDPM without guidance and Ours (icDPM with guidance).
Figure 15. **HIDE** [64] deblurring examples from MPRNet [82], MIMO UNet+ [5], SAPHNet [71], Restormer [80], UFormer [76], DvSR [77] and Ours.
Figure 16. HIDE [64] deblurring examples from MPRNet [82], MIMO UNet+ [5], SAPHNet [71], Restormer [80], UFormer [76], DvSR [77] and Ours.
Figure 17. **HIDE [64]** deblurring examples from MPRNet [82], MIMO UNet+ [5], SAPHNet [71], Restormer [80], UFormer [76], DvSR [77] and Ours.
Figure 18. **HIDE** [62] deblurring examples from MPRNet [82], MIMO UNet+ [5], SAPHNet [71], Restormer [80], UFormer [76], DvSR [77] and Ours.
Figure 19. GoPro [49] deblurring examples from MPRNet [82], MIMO UNet+ [5], SAPHNet [71], Restormer [80], UFormer [76], DvSR [77] and Ours.
Figure 20. GoPro [49] deblurring examples from MPRNet [82], MIMO UNet+ [5], SAPHNet [71], Restormer [80], UFormer [76], DvSR [77] and Ours.
Figure 21. GoPro [49] deblurring examples from MPRNet [82], MIMO UNet+ [5], SAPHNet [71], Restormer [80], UFormer [76], DvSR [77] and Ours.
Figure 22. **GoPro** [49] deblurring examples from MPRNet [82], MIMO UNet+ [5], SAPHNet [71], Restormer [80], UFormer [76], DvSR [77] and Ours.
Figure 23. Realblur-J [56] deblurring examples from UNet [59], MPRNet [82], DeblurGAN-v2 [32], Restormer [80], UFormer [76], DvSR [77] and Ours.
Figure 24. **Realblur-J** [56] deblurring examples from UNet [59], MPRNet [82], DeblurGAN-v2 [32], Restormer [80], UFormer [76], DvSR [77] and Ours.
Figure 25. **Realblur-J** [56] deblurring examples from UNet [59], MPRNet [82], DeblurGAN-v2 [32], Restormer [80], UFormer [76], DvSR [77] and Ours.
Figure 26. **Realblur-J** [56] deblurring examples from UNet [59], MPRNet [82], DeblurGAN-v2 [32], Restormer [80], UFormer [76], DvSR [77] and Ours.
Figure 27. Failure case from Realblur-J [56] in low-light scenes.
Figure 28. Failure case from Realblur-J [56] with strong light streaks.
Figure 29. Failure case from Realblur-J [56] in night scenes.
Figure 30. Failure case from Realblur-J [56] in low-light condition.