Residual Diffusion Deblurring Model for Single Image Defocus Deblurring

Haoxuan Feng[1], Haohui Zhou[1], Tian Ye[1], Sixiang Chen[1], Lei Zhu[1,2*]

[1]The Hong Kong University of Science and Technology (Guangzhou)
[2]The Hong Kong University of Science and Technology
{hfeng108, hzhou142, tye610, schen691}@connect.hkust-gz.edu.cn,
leizhu@ust.hk

Abstract

Defocus deblurring is a challenging task due to the spatially varying nature of defocus blur with multiple plausible solutions of a single given image. However, most existing methods falter when faced with extensive and variable defocus blur, either ignoring it or relying on additional loss functions to enhance perceptual quality. This often results in unrealistic reconstructions and compromised generalizability. In this paper, we propose a novel Residual Diffusion Deblurring Model framework for single image defocus deblurring. Our approach integrates a pre-trained defocus map estimator and a lightweight pre-deblur module with a learnable receptive field, providing crucial posterior information to effectively address large-scale and varying shaped defocus blur. In addition, a carefully-design denoising network enables the generation of diverse reconstructions from a single input. This approach not only significantly improves the perceptual quality of defocus deblurring outputs through multi-step residual learning, but also offers a more efficient inference strategy. Experimental results demonstrate that our method achieves competitive performance on real-world defocus deblurring image datasets across both perceptual and distortion evaluation metrics.

Introduction

Defocus deblurring is an important task in image processing and computer vision that aims to restore a all-in-focus image from a defocused image (Abuolaim and Brown 2020; Ruan et al. 2022; Cun and Pun 2020). Defocus blur is always caused by using a large aperture to increase the luminous flux and thus shorten the exposure time to capture an image. This creates a shallow depth of field (DoF), causing points far from the focal plane to be projected onto the camera sensor as out-of-focus circles of confusion (COC) that exceed the permitted diameter (Potmesil and Chakravarty 1981). Using a smaller aperture will provide a deeper depth of field but reducing the vision quality and increasing the exposure time will make it easier to produce motion blur on moving objects. The complexity of defocus blurring stems from its spatial variability and the fact that each pixel point produces a different COC diameter, which significantly undermines image visual quality by obfuscating crucial details and information. This degradation can adversely impact downstream computer vision tasks, including object detection (He et al. 2024a) and semantic segmentation (Xie et al. 2021), making the defocus deblurring an important challenge.

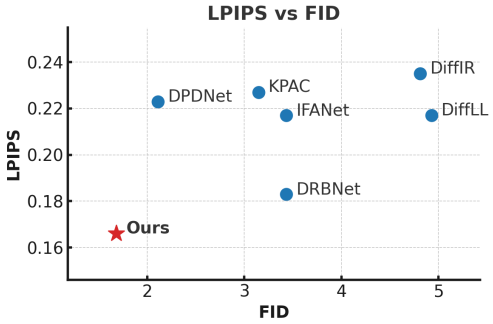

Figure 1: Perceptual evaluation metrics (LPIPS and FID_{192}) comparison on DPDD dataset (Abuolaim and Brown 2020) demonstrates that our RDDM achieves the state-of-the-art performance on single image defocus deblurring task.

Traditional approaches (Zhuo and Sim 2011; Karaali and Jung 2018; Levin et al. 2007; D'Andres et al. 2016; Lee et al. 2019) to defocus blur mitigation typically employ a two-step procedure: first estimating the per-pixel blur kernel, followed by applying non-blind deconvolution to obtain an all-in-focus image. Although these methods are simple and effective for addressing uniform blur, they often struggle with the inherent non-linearity and varying shapes of real-world defocus blur, which defy confinement to predefined shapes such as disc or simple Gaussian kernels. Furthermore, even with an accurately estimated kernel, the Gibbs phenomenon (Yuan et al. 2007) introduces ringing artefacts that present additional challenges.

In recent years, deep learning strategies have attempted to directly restore sharp images from blurred inputs by leveraging the modeling capabilities of neural networks to adapt to various blur distributions (Lee et al. 2021; Cun and Pun 2020; Abuolaim and Brown 2020; Ruan et al. 2022; He et al. 2024b; Ye et al. 2022). These methods typically optimize pixel-level losses (e.g., L1, L2) to improve metrics such as PSNR. However, two significant shortcomings re-

*Lei Zhu (leizhu@ust.hk) is the corresponding author.

main: 1) while technically accurate, the results frequently **lack promising perceptual details**, leading to dissatisfaction in practical applications; 2) they fail to effectively manage **large and complex defocus blurs**, as evidenced by persistent issues in both robustness and adaptability. Existing methods add additional loss terms (Delbracio, Talebei, and Milanfar 2021; Mechrez, Talmi, and Zelnik-Manor 2018; Zhai et al. 2023) to improve the generated image quality based on human perception, but these strategies always suffer from overfitting problem on the training distribution and fail with generalisation ability to handle out-of-distribution data samples. Also, (Abuolaim and Brown 2020; Lee et al. 2021) employing larger receptive fields and spatially-varying per-pixel deblurring filters aim to address large defocus blur challenges. However, these approaches often cannot sufficiently capture the complex, spatially varying blur encountered in real-world images, leading to suboptimal deblurring performance for both perceptual and distortion metrics.

To address the challenges mentioned previously, the defocus deblurring task can be conceptualized as a conditional generative task. In this context, the objective is to identify the optimal sample from a specified posterior distribution through the application of various hyperparameters. More specifically, we propose a **Residual Diffusion Deblurring Model (RDDM)** that integrates a conditional diffusion model guided by a defocus map and incorporates residual learning into the single image defocus deblurring process. To effectively and accurately handle large defocus blur, which necessitates expansive receptive fields, our architecture first approximates the blur shape using a preliminary, low-cost estimate provided by a lightweight pre-deblur module guided by a pre-trained defocus map estimator. Then, we utilize a carefully designed deblurring network that uses residual learning to perform detail refinements step-by-step. Our denoising network in RDDM is powered by learnable receptive fields, which significantly improves the perceptual quality of the deblurred images. From what we know, RDDM is the first diffusion method designed specifically for the single image defocus deblurring task. We aim for our work to inspire further advancements in this field.

Overall, our contributions can be summarized as follows:

- We propose a novel diffusion-based defocus deblurring framework equipped with a light-weight pre-deblur module and a carefully-designed deblurring module to generate realistic details for only 2$\sim$4 steps based on the initial pre-deblurred result.
- To effectively incorporate the information from the guidance and handle large defocus blur, we propose **Residual Diffusion Deblur** (RDD) block, which combines the spatially varying information and learnable receptive field with rich semantic guidance.
- We conduct extensive experiments on DPDD dataset and demonstrate the state-of-the-art performance on it and enhance the perceptual quality significantly. Also, we evaluate our generalization ability on real-world out-of-distribution datasets such as RealDOF dataset and CUHK dataset.

Related Work

Conventional Methods

Traditional defocus deblurring methods (Zhuo and Sim 2011; Karaali and Jung 2018; Levin et al. 2007; D'Andres et al. 2016; Lee et al. 2019) predominantly utilize handcrafted features and employ a two-stage process. Initially, they estimate the defocus blur for each pixel, subsequently followed by non-blind deconvolution. However, these methods struggle with real-world blurring phenomena, characterized by non-linearity and variable shapes. Consequently, Gaussian or linear kernels fall short in precisely modeling real-world defocus blur.

Deep-learning Based Methods

Recently, CNN-based methods (Whang et al. 2022; Lee et al. 2021; Cun and Pun 2020; Abuolaim and Brown 2020; Ruan et al. 2022) have significantly enhanced a variety of vision applications, including defocus deblurring. Early two steps method (Lee et al. 2019) follows the deblurring strategy based on conventional methods, which employ an end-to-end defocus map estimating network to predict an accurate blur kernel and apply blind/non-blind deconvolution algorithm to gain the deblurred image. However, estimating an accurate blur kernel is hard and even accurate kernel may also not reach a competitive performance. Base on this challenge, DPDNet propose (Abuolaim and Brown 2020) an end-to-end learning-based methodology to surpasses traditional approaches. Nonetheless, it struggles with handling spatially varying and large defocus blur. To address this, IFANet (Lee et al. 2021) introduces spatially-adaptive per-pixel deblurring filters applied to the blurred image, generating deblur features rather than directly predicting pixel-level values. However, its training and inference stages are distinct, failing to account for misalignment in training pairs. (Ruan et al. 2022) suggests a pretrain-finetune strategy to tackle inconsistencies in DPDD, even in sharp regions, using pretrained models on LFDOF (Ruan et al. 2021). Moreover, generative models like DefocusGAN (Zhai et al. 2023) significantly improve perceptual quality through adversarial loss and unsupervised learning strategies, albeit struggling with unexpected artifacts absent in the source clean image.

Methodology

DDPM and DDIM

DDPM Before introducing our method, some concepts about understanding diffusion process should be reviewed (Luo 2022). Diffusion probabilistic models (DPM) (Ho, Jain, and Abbeel 2020; Sohl-Dickstein et al. 2015), is a T-step Markov chain $(x_0, \ldots, x_T)$ that starts from a ground truth image sampled from the real world data distribution $x_0 \in R^d$ and repeatedly adds Gaussian noise until pure random Gaussian noise z. This process called the forward process and can be summarized as follows:

$$p(x_t|x_{t-1}) = N(x_t; \sqrt{\alpha_t}x_{t-1}, (1-\alpha_t)I_d), \qquad (1)$$

Or Equivalently:

$$x_t = \sqrt{\alpha_t} x_{t-1} + \sqrt{1-\alpha_t}\epsilon_t, \tag{2}$$

Where α_t is the noise level controller and judged as the hyperparameter during forward step, which represent the variance of the noise addition level of each step. Then, since equation 1 is only one step of adding noise and all ϵ_t is i.i.d for each step, based on the normal distribution addition formula, we can summarize the whole from x_{t-1} to x_0:

$$x_t = \sqrt{\prod_{i=1}^{t} \alpha_i} x_0 + \sqrt{1-\prod \alpha_i}\epsilon_0, \epsilon \in N(0, I_d), \tag{3}$$

Let $\prod_{i=1}^{t} \alpha_i = \overline{\alpha_t}$, the formula can be simplified to $x_t = \sqrt{\overline{\alpha_t}} x_0 + \sqrt{1-\overline{\alpha_t}}\epsilon$ This process reduces the forward step (adding noise) and allow us to add noise directly from sampling ground truth image x_0 to random noise z. Based on previous strategy, DDPM aims to learned a reverse process which can recover the ground truth image from a random noise input z, which is $p(x_{t-1}|x_t)$. From bayes rules, the reverse step is:

$$p(x_{t-1}|x_t, x_0) = N(x_{t-1}; u(x_t, x_0), \beta_t I_d), \tag{4}$$

where $u(x_t, x_0) = \frac{\sqrt{\alpha_t}(1-\overline{\alpha}_{t-1})x_t + \sqrt{\overline{\alpha}_{t-1}}(1-\alpha_t)x_0}{1-\overline{\alpha_t}}$ and $\beta_t = \frac{(1-\alpha_t)(1-\overline{\alpha}_{t-1})}{1-\overline{\alpha}_{t-1}}$. Since the reverse denoising step does not condition x_0, this process will be parameterized by a neural network $f_\theta(x_t, t)$ which takes noisy image x_t and timesteps t as input and predict clean image x_0, the parameterized function is:

$$u_\theta(x_t, t) = \frac{\sqrt{\alpha_t}(1-\overline{\alpha}_{t-1})x_t + \sqrt{\overline{\alpha}_{t-1}}(1-\alpha_t)f_\theta(x_t, t)}{1-\overline{\alpha_t}}, \tag{5}$$

Finally, the learning objective of this function is to minimize the KL divergence of the ground truth denoising process $p(x_{t-1}|x_t, x_0)$ and the estimating $p_\theta(x_{t-1}|x_t)$, which can be equivalently to minimize the distance between ground truth image x_0 with estimation image $x_\theta(x_t, t)$. In our experiment, we utilized PSNR Loss as (Chen et al. 2022) do as the optimization function.

DDIM DDPM (Ho, Jain, and Abbeel 2020) suffer from the inference speed which related to the Markov chain strategy and cannot predict samples based on non-neighbor point. Several methods (Lee et al. 2022; Song, Meng, and Ermon 2022; Kong and Ping 2021; Jolicoeur-Martineau et al. 2021) aim to reduce the inference time here. DDIM (Song, Meng, and Ermon 2022) solve this problem by remove the closed form solution of $p(x_t|x_{t-1})$ because the optimize function and the sampling strategy is only correlated to $p(x_t|x_0)$ and $p(x_{t-1}|x_t)$. Assume that:

$$p(x_{t-1}|x_t, x_0) = N(x_{t-1}; k_t x_0 + m_t x_t, \sigma_t^2 I), \tag{6}$$

By applying Bayes rules, the result should only satisfy:

$$\int p(x_{t-1}|x_t, x_0)p(x_t|x_0)dx_t = p(x_{t-1}|x_0), \tag{7}$$

Algorithm 1: Inference procedure of our methods

Require: Pre-deblur module p_θ, defocus map estimator g_θ, denoising network f_θ, blurred image I_b, num of training timestep T_l and num of inference timestep T_i
Ensure: Deblurred image I

```
I_p = p_θ(I_b)                          ▷ Got Pre-deblurred image
I_map = g_θ(I_p)                        ▷ Got defocus map
x_t ~ N(0, I_d)                         ▷ Sampling noise from gaussian
for t in range (T_l, 1, T_l/T_i) do
    z_t ~ N(0, I_d)
    x_{t-1} = m_t x_t + k_t f_θ(x_t, t, c) + σ β_t z_t
                                        ▷ Reverse diffusion process
end for
return I_p + x_{t-1}                    ▷ Final deblurred image
```

With the method of undermined coefficient, we can solve the following two equations to get the ground truth reverse process by:

$$\begin{cases} \sqrt{\overline{\alpha}_{t-1}} = m_t \overline{\alpha}_t + k_t, \\ m_t^2(1-\overline{\alpha}_t) + \sigma^2 = 1 - \overline{\alpha}_{t-1} \end{cases} \tag{8}$$

So, the coefficient solutions are $m_t = \frac{\sqrt{1-\overline{\alpha}_{t-1}-\sigma^2}}{1-\overline{\alpha}_{t-1}}$ and $k_t = \sqrt{\overline{\alpha}_{t-1}} - \frac{\sqrt{1-\overline{\alpha}_{t-1}-\sigma^2}}{1-\overline{\alpha}_{t-1}}\sqrt{\overline{\alpha}_t}$. Based on this solution, the ground truth estimation function can be calculated correctly. Also, the training process is the same as DDPM, where we estimate a denoising function $f_\theta(x_t, t)$ to estimate the result and utilize distance-based as loss function to optimize the model. However, unlike DDPM, DDIM has another hyperparameter σ to estimate the variance of denoising process and in our practice the σ is set to 0. In conclusion, the training process of DDPM and DDIM is the same and the difference between them are during inference process where DDIM only need a few steps to generate result.

Conditional Diffusion In order to control the denoising process and reduce the generation variance, some methods utilize (Lee et al. 2022; Saharia et al. 2022) class as guidance and in our task, the iteration process is dependent on both time embedding t and the condition y. The training objective is:

$$L_\theta = E||f_\theta(x_t, t, y) - x_0||_p^p, \tag{9}$$

where f_θ, x_0, p is the denoising function, ground truth image and p-norm value.

Preliminary

Single image defocus deblurring aims to restore the all-in-focus image $\hat{I}$ from an given blurred image I_b. By parameterized a deep learning network f_θ and given some guidance input c, the whole process can be summarized as:

$$\hat{I} = f_\theta(I_b, c), \tag{10}$$

As illustrated in Figure 2, for a blurred image input $I_b \in R^{H\times W\times 3}$, a pretrained defocus map estimator g_θ first predicts a defocus map $I_{map} \in R^{H\times W\times 1}$ that serves as guidance. Subsequently, a pre-deblur module p_θ processes both

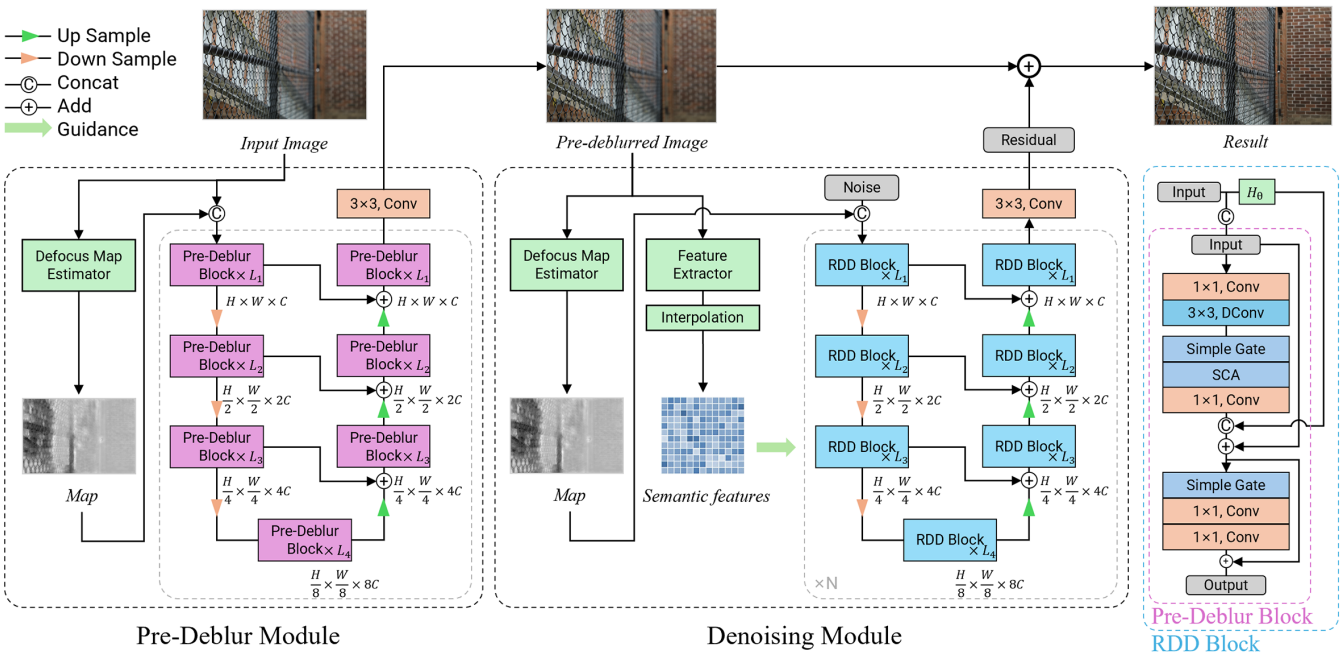

Figure 2: The overview our RDDM framework. Given an input image, the **Pre-deblur-module** will take both blurred image and defocus map to give a cheap guess output. Then, the **defocus map estimator** will evaluate an accurate map based on the Pre-deblur module and concat with pre-deblurred image as guidance of **denoising module**. Finally, after passing several foward times, the output residual will add with pre-deblurred image to gain the reconstruction image.

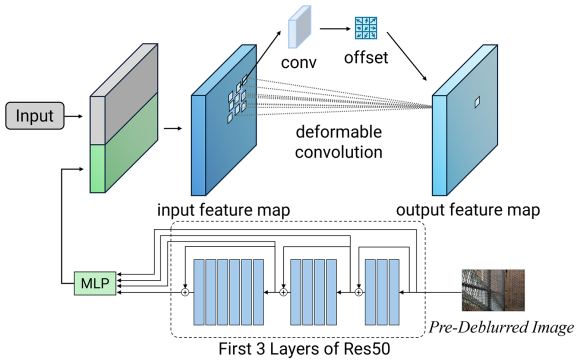

Figure 3: Details of the RDD input, which includes a guidance network (He et al. 2015) and a deformable convolution layer (Xiong et al. 2024).

the blurred image I_b and the defocus map I_{map} to produce a pre-deblurred image I_p. This pre-deblurred image is then concatenated with the defocus map to serve as input for the denoising network f_θ, which executes N inference steps to produce the predicted all-in-focus image. The complete inference process is detailed in Algorithm 1.

Pre-deblur Block and RDD Block

Given that NAFNet (Chen et al. 2022) has established itself as a baseline in image restoration tasks, we select it as our backbone model. To address both inter-complexity and intra-complexity in NAFNet, we employed several strategies, including simplified channel attention. However, its basic structural design limits performance, particularly in accurately focusing on blurred positions within our task. To overcome this limitation, we propose the Pre-deblur Block, which substitutes the depth-wise convolution layer in a standard NAFBlock with a Deformable convolution (Xiong et al. 2024), coupled with Layer Normalization (Ba, Kiros, and Hinton 2016) across a larger receptive field to more effectively manage defocus blur. Furthermore, we propose **Residual diffusion deblur (RDD)** block to and tackle large-scale defocus blur and enhance detail refinement with additionally add a semantic guidance network H_θ based on pre-deblur block. Each RDD block processes both the output from the previous block $\hat{x}$ and a semantic guidance feature s. The process begins with a combination of 1×1 Convolution and 3×3 deformable convolution to establish a learnable receptive field. Following (Chen et al. 2022), a simplified channel attention mechanism is integrated into our block to capture global information efficiently. Additionally, following (Ren et al. 2023), we concatenate the semantic guidance network output after channel attention layer with additional multi-scaled guidance pass through guidance network for extracting the coarse structure features from the input at multiple resolutions. The whole RDD block with downsampling layer can be summarized as follows:

$$\hat{y}_\downarrow = RDD(\hat{x}_\downarrow, H_\theta(\hat{x}_\downarrow)), \tag{11}$$

To optimize the inference steps and ensure that the denoising process primarily focuses on deblurring rather than on broader image generation, we have chosen to implement residual learning (Whang et al. 2022). This technique allows the denoising network to model the residuals based on a preliminary estimation provided by the Pre-deblur module. This approach shifts the network's focus from reconstructing an entire image to refining a rough initial guess into a clearer output. Consequently, the overall training objective for our system can be outlined as follows:

$$L_\theta = E||(x_0 - p_\theta(I_b, I_{map})) - f_\theta(x_t, t, c)||_1, \quad (12)$$

where p_θ is the Pre-deblur module and f_θ is the denoising network base some condition c. We introduce the pseudocode in Algorithm 1 and the Pre-deblur module not need extra loss to optimize the performance of it. Notice that the Pre-deblur module is construct based on the Pre-deblur block mention previously. Specifically, we downsampled the image resolution to $\frac{1}{8}$ and connect each Pre-deblur block follows U-Net structure (Abuolaim and Brown 2020; Chen et al. 2022). Contrary to the approach described in (Whang et al. 2022), our experimental results indicated that using a larger Pre-deblur module coupled with a smaller denoising network did not yield satisfactory outcomes for our specific task. Consequently, we adopted the opposite configuration: a simpler, cost-effective Pre-deblur module coupled with a more extensive denoising network. Also, since blurred image will lead to misunderstand artifacts as image content and further affect the final performance, Pre-deblur module provides a more suitable condition for denoising network. This configuration allows us to efficiently model the residuals based on the initial estimations made by the Pre-deblur module. This strategic choice proves effective in handling large-scale defocus blur at high resolutions, enabling the denoising process to achieve well-reconstructed results within only 2 to 4 inference steps. Regarding the Defocus map estimator, we employed a SegFormer (Xie et al. 2021), following the methodology of (Lee et al. 2019), and trained it on a real-world defocus blur dataset. Specifically, SegFormer model was trained on the dataset provided by (Lee et al. 2019), which comprised a set of real defocused images with ground truth binary blur map. The loss function was set to a sigmoid cross-entropy loss and Adam (Loshchilov and Hutter 2017) was used for optimization. For a deeper understanding of the defocus map generation and its application, we recommend that readers consult (Lee et al. 2019).

Denoising Network of RDDM

The denoising network is constructed based on the Residual Diffusion Deblur (RDD) block, which is integrated at each scale of the corresponding decoder within a U-Net architecture. This integration facilitates a progressive restoration of the all-in-focus image. Specifically, both the encoder and the decoder comprise four blocks. Each encoder block includes an additional downsampling convolution, while each decoder block utilizes a pixel shuffle strategy (Shi et al. 2016) to upscale the feature maps efficiently. Also, as shown in Figure 2, we use ResNet50 (He et al. 2015) as backbone and choose the first 3 layer of it as the semantic information of blurred input. The guidance network H_θ is constructed with two 1×1 convolution layer with LeakyReLU (Xu et al. 2015) activation function with bilinear interpolation for different resolution to provide sufficient multi-scale semantic guidance for each RDD block. The whole denoising network can be summarized as follows:

$$c = \textbf{Concat}(I_p, H_\theta(Res50(I_p))), \quad (13)$$

$$R_t = f_\theta(x_t, t, c), \quad (14)$$

Where s, c, r_t represent semantic guidance, denoising guidance and predict residual at time step t. Finally, after n inference steps, the final reconstruction image I becomes:

$$I = I_p + R_n, \quad (15)$$

Dataset	#Image Resolution, Source
CUHK (2014)	$\sim 470 \times 610$, Internet
DPDD (2020)	1120×1680, Canon 5D Mark IV
RealDOF (2021)	$\sim 1536 \times 2320$, Sony $\alpha 7R$ IV

Table 1: Defocus Blur Datasets statistics

Experiments

Dataset We train our methods on **DPDD** dataset (Abuolaim and Brown 2020), which includes 2,000 images distributed across 500 scenes. Each scene features a defocus blur captured with a large aperture alongside a corresponding all-in-focus image using a small aperture. The dataset also contains two associated DP sub-aperture views, however, we don't use this information as our study focuses exclusively on single image deblurring. To assess the generalization capabilities of our models, we also tested them on the **CUHK** dataset (Shi, Xu, and Jia 2014) which consists of 1,000 blurred images from the Internet, including both motion and defocus blurs, though our focus was only on the latter. Additionally, we evaluated our models on the **RealDOF** (Lee et al. 2021) dataset, which comprises 50 scenes captured with Sony a7R IV cameras. Detailed dataset statistics are available in Table 1.

Implementation Details We implement our models using PyTorch (Paszke et al. 2019). We use AdamW (Loshchilov and Hutter 2017) with $\beta_1 = 0.9$ and $\beta_2 = 0.999$ and weight decay to 0.01 for training. The network is trained for 150k steps with an initial learning rate of $2e-4$. Also, the cosine learning rate scheduler is to enhance training stability, returning to the initial learning rate after 200 step updates. For the diffusion hyperparameters, we choose DDIM (Song, Meng, and Ermon 2022) sampler and set the number of training steps and inference steps to 1,300 and 3, respectively. Data augmentation and loading involve using a crop size of 320×320 with a batch size of 8, and augment the data with horizontal flipping, vertical flipping and random rotation. For evaluation of diffusion deblurring performance, we measure the PSNR, SSIM (Wang et al. 2004), LPIPS (Zhang et al. 2018) and FID_{192} (Heusel et al. 2018) between blurred image and ground truth images on a single Nvidia A800-SXM4 GPU with a batch size of 1, indicating that the experiment is not affected by batch-dependent noise.

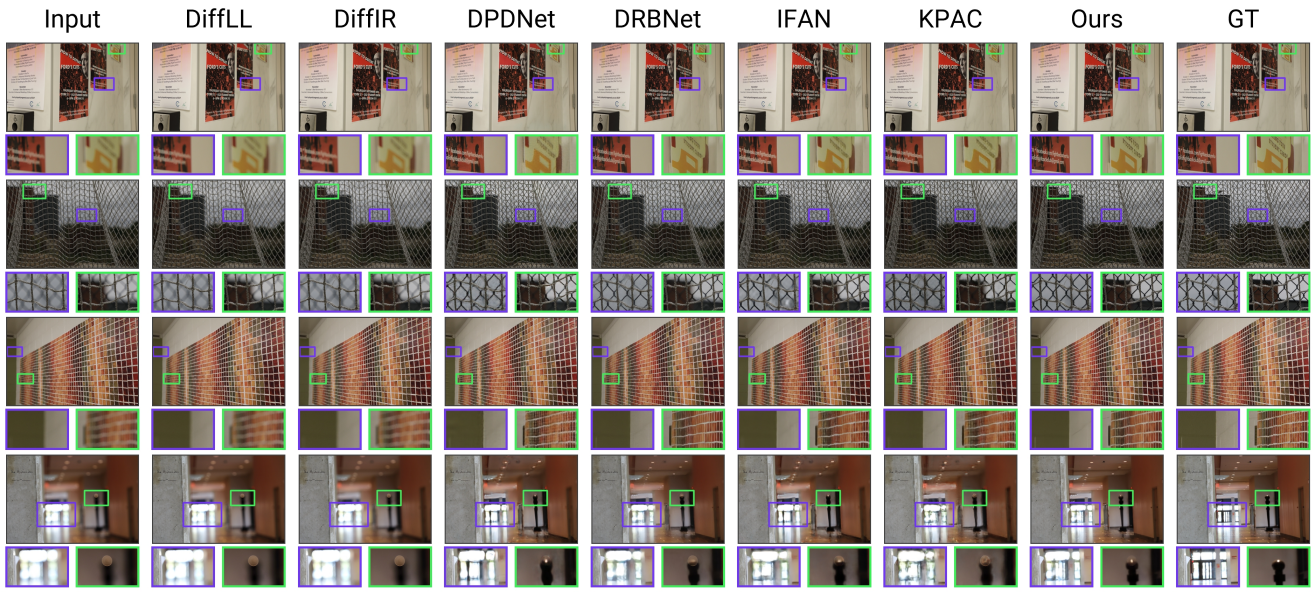

Figure 4: Experiment results on **DPDD** (Abuolaim and Brown 2020) dataset among **DPDNet** (Abuolaim and Brown 2020), **KPAC** (Son et al. 2021), **IFANet** (Lee et al. 2021) **DRBNet** (Ruan et al. 2022), **DiffLL** (Jiang et al. 2023), **DiffIR** (Xia et al. 2023) and **ours**.

Comparison of Experimental Result

We compare our methods with current end-to-end training-based approaches such as DPDNet (Abuolaim and Brown 2020), KPAC (Son et al. 2021), IFANs (Lee et al. 2021), and DRBNet (Ruan et al. 2022) and diffusion-based image restoration approaches DiffIR (Xia et al. 2023) and DiffLL (Jiang et al. 2023). It is noteworthy that DiffIR and DiffLL are not designed for defocus deblurring specifically; therefore, they are retrained on the DPDD dataset (Abuolaim and Brown 2020) (implementation details can be found in the supplementary material). In contrast, the remaining methods are end-to-end learning-based approaches that directly recover the clear image from blurred inputs. All mentioned methods use performance metrics proposed in the original papers and utilize parameters provided by the authors.

The quantitative evaluation on the DPDD test dataset is shown on Table 2. For end-to-end learning-based models, our approach outperforms the current best method, DRBNet, by 9.2% (0.017) and exceeds other methods such as IFANs by 23.5% (0.051), KPAC by 26.8% (0.06) and DPDNet by 25.5% (0.057) based on LPIPS perceptual evaluation metric respectively. As mentioned previously, DPDNet (Notice that here we report the dual pixel result for DPDNet$_D$ since the performance better than single image deblurring and the author do not release the pretrained parameters of DPDNet$_S$) shows poorer performance on end-to-end training models due to the naive U-Net architecture and the random artifacts it produced. However, dual-pixel view can help to better refine the details of image which shows performance in perceptual quality. KPAC perform better than DPDNet but the performance is restricted based on the small training parameters. Both IFANs and DRBNet achieve competitive performance, but IFAN also needs dual pixel view to gain disparity map and DRBNet need LFDOF dataset for a pretrain-finetune strategy. For diffusion-based image restoration method, our approach significantly outperforms both DiffLL and DiffIR by 32.7% and 29.3% based on LPIPS perceptual evaluation metric respectively. Meanwhile, our method excels in all evaluation metrics and only requires a single view from the DPDD dataset. Figure 4 illustrates the perceptual quality of all models mentioned above.

Generalization Ability

To assess the generalization ability of our model, we utilize two datasets: 1) The CUHK blur detection dataset (Shi, Xu, and Jia 2014), which includes 704 defocus images with a resolution of 470×610, collected from the Internet and lacking all-in-focus ground truth images. 2) The RealDOF dataset (Lee et al. 2021), consisting of 50 scenes with corresponding all-in-focus images captured by Sony a7R IV cameras. The results demonstrate that our method more effectively removes defocus blur and refines details better than all others methods and maintains competitive visual quality in the RealDOF dataset. To show the numerical comparison on generalization ability, we also compute the perceptual and distortion quality on RealDOF dataset, which contains a lot of large defocus blur scenes on out-door distribution. For a fair comparison, the methods listed in Table 2 are trained solely on the DPDD dataset and evaluated on the RealDOF

Model	DPDD Dataset					RealDOF Dataset				
	LPIPS↓	MAE↓	FID↓	PSNR↑	SSIM↑	LPIPS↓	MAE↓	FID↓	PSNR↑	SSIM↑
DPDNet (2020)	0.223	0.041	2.112	25.13	0.786	0.402	0.042	6.662	24.48	0.744
KPAC (2021)	0.227	0.040	3.146	25.22	0.774	0.349	0.055	6.421	23.34	0.747
IFANet (2021)	0.217	0.039	2.355	25.37	0.789	0.304	0.041	**2.706**	24.71	0.763
DRBNet (2022)	0.183	0.039	3.433	25.73	0.791	0.329	**0.036**	3.104	24.68	0.751
DiffLL (2023)	0.247	0.045	4.928	23.52	0.745	0.489	0.061	18.25	22.58	0.682
DiffIR (2023)	0.235	0.043	4.812	23.97	0.762	0.423	0.052	13.10	23.09	0.706
RDDM (Ours)	**0.166**	**0.037**	**1.681**	**25.97**	**0.811**	**0.271**	0.038	2.912	**25.03**	**0.772**

Table 2: Defocus deblurring performance on **DPDD** (Abuolaim and Brown 2020) and **RealDOF** (Lee et al. 2021) dataset. In the case of the diffusion-based method, the numbers of inference steps were set to 4 (**DiffIR** (Xia et al. 2023)), 10 (**DiffLL** (Jiang et al. 2023)) and 2 (**Ours**).

Figure 5: Perceptual quality results on **RealDOF** (Lee et al. 2021) dataset among **DPDNet** (Abuolaim and Brown 2020), **KPAC** (Son et al. 2021), **IFANet** (Lee et al. 2021), **DRBNet** (Ruan et al. 2022), **DiffLL** (Jiang et al. 2023), **DiffIR** (Xia et al. 2023) and **Ours**.

test set. Table 2 shows the quantitative evaluation on RealDOF dataset. Our method exhibits state-of-the-art performance in all distortion-based evaluation metrics and demonstrates significantly competitive performance in perceptual quality. Specifically, for perceptual metrics like LPIPS and FID, our method outperforms all other deep-learning-based methods by 10.8% (0.033) compared to IFANet, the second-best method, and achieves nearly competitive performance on FID, slightly trailing by 7.1%. As previously mentioned, the RealDOF dataset contains numerous scenes with large defocus blur. Figure 5 illustrates the qualitative results, showcasing our method's focus on refining details such as patterns and text. Detail analysis include ablation study and analysis on model design are provided in the supplementary material.

Conclusion

In this paper, we introduce a novel defocus deblurring framework based on a conditional diffusion model, capable of capturing spatially-varying information with a learnable receptive field to address real-world defocus deblurring challenges. To efficiently and accurately handle large and varying defocus blur, we propose a predeblur module equipped with a pretrained defocus map estimator. For enhanced perceptual quality and precise detail reconstruction, we have developed a RDD block that constructs the denoising network, supplemented by an additional semantic guidance network for iterative refinement. Extensive experimental results on real-world defocus blur datasets demonstrate that our method significantly improves perceptual quality and also achieves competitive performance in distortion-based evaluation metrics with robust generalization capabilities.

Acknowledgments

This work is supported by the Guangzhou-HKUST(GZ) Joint Funding Program (No. 2023A03J0671), the Guangzhou Municipal Science and Technology Project (Grant No. 2024A04J4230), the Guangzhou Industrial Information and Intelligent Key Laboratory Project (No. 2024A03J0628), the Nansha Key Area Science and Technology Project (No. 2023ZD003), Guangdong Provincial Key Lab of Integrated Communication, Sensing and Computation for Ubiquitous Internet of Things(No.2023B1212010007).

References

Abuolaim, A.; and Brown, M. S. 2020. Defocus Deblurring Using Dual-Pixel Data. arxiv:2005.00305.

Ba, J.; Kiros, J. R.; and Hinton, G. E. 2016. Layer Normalization. *ArXiv*, abs/1607.06450.

Chen, L.; Chu, X.; Zhang, X.; and Sun, J. 2022. Simple Baselines for Image Restoration. arxiv:2204.04676.

Cun, X.; and Pun, C.-M. 2020. Defocus Blur Detection via Depth Distillation. *ArXiv*, abs/2007.08113.

D'Andres, L.; Salvador, J.; Kochale, A.; and Susstrunk, S. 2016. Non-Parametric Blur Map Regression for Depth of Field Extension. *IEEE transactions on image processing: a publication of the IEEE Signal Processing Society*, 25(4): 1660–1673.

Delbracio, M.; Talebei, H.; and Milanfar, P. 2021. Projected Distribution Loss for Image Enhancement. In *2021 IEEE International Conference on Computational Photography (ICCP)*, 1–12.

He, C.; Li, K.; Zhang, Y.; Xu, G.; Tang, L.; Zhang, Y.; Guo, Z.; and Li, X. 2024a. Weakly-supervised concealed object segmentation with sam-based pseudo labeling and multi-scale feature grouping. *Advances in Neural Information Processing Systems*, 36.

He, C.; Shen, Y.; Fang, C.; Xiao, F.; Tang, L.; Zhang, Y.; Zuo, W.; Guo, Z. H.; and Li, X. 2024b. Diffusion Models in Low-Level Vision: A Survey. *ArXiv*, abs/2406.11138.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2015. Deep Residual Learning for Image Recognition. arxiv:1512.03385.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2018. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. arxiv:1706.08500.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising Diffusion Probabilistic Models. arxiv:2006.11239.

Jiang, H.; Luo, A.; Han, S.; Fan, H.; and Liu, S. 2023. Low-Light Image Enhancement with Wavelet-Based Diffusion Models. *ACM Transactions on Graphics (TOG)*, 42: 1 – 14.

Jolicoeur-Martineau, A.; Li, K.; Piché-Taillefer, R.; Kachman, T.; and Mitliagkas, I. 2021. Gotta Go Fast When Generating Data with Score-Based Models. arxiv:2105.14080.

Karaali, A.; and Jung, C. R. 2018. Edge-Based Defocus Blur Estimation With Adaptive Scale Selection. *IEEE Transactions on Image Processing*, 27(3): 1126–1137.

Kong, Z.; and Ping, W. 2021. On Fast Sampling of Diffusion Probabilistic Models. arxiv:2106.00132.

Lee, J.; Lee, S.; Cho, S.; and Lee, S. 2019. Deep Defocus Map Estimation Using Domain Adaptation. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 12214–12222. Long Beach, CA, USA: IEEE. ISBN 978-1-72813-293-8.

Lee, J.; Son, H.; Rim, J.; Cho, S.; and Lee, S. 2021. Iterative Filter Adaptive Network for Single Image Defocus Deblurring. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2034–2042. Nashville, TN, USA: IEEE. ISBN 978-1-66544-509-2.

Lee, S.-g.; Kim, H.; Shin, C.; Tan, X.; Liu, C.; Meng, Q.; Qin, T.; Chen, W.; Yoon, S.; and Liu, T.-Y. 2022. PriorGrad: Improving Conditional Denoising Diffusion Models with Data-Dependent Adaptive Prior. arxiv:2106.06406.

Levin, A.; Fergus, R.; Durand, F.; and Freeman, W. T. 2007. Image and Depth from a Conventional Camera with a Coded Aperture. *ACM Transactions on Graphics*, 26(3): 70–es.

Loshchilov, I.; and Hutter, F. 2017. Fixing Weight Decay Regularization in Adam. *ArXiv*, abs/1711.05101.

Luo, C. 2022. Understanding Diffusion Models: A Unified Perspective. arxiv:2208.11970.

Mechrez, R.; Talmi, I.; and Zelnik-Manor, L. 2018. The Contextual Loss for Image Transformation with Non-Aligned Data. *ArXiv*, abs/1803.02077.

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; Desmaison, A.; Köpf, A.; Yang, E.; DeVito, Z.; Raison, M.; Tejani, A.; Chilamkurthy, S.; Steiner, B.; Fang, L.; Bai, J.; and Chintala, S. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. *ArXiv*, abs/1912.01703.

Potmesil, M.; and Chakravarty, I. 1981. A Lens and Aperture Camera Model for Synthetic Image Generation. *ACM SIGGRAPH Computer Graphics*, 15(3): 297–305.

Ren, M.; Delbracio, M.; Talebi, H.; Gerig, G.; and Milanfar, P. 2023. Multiscale Structure Guided Diffusion for Image Deblurring. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 10721–10733.

Ruan, L.; Chen, B.; Li, J.; and Lam, M. 2022. Learning to Deblur Using Light Field Generated and Real Defocus Images. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 16283–16292. New Orleans, LA, USA: IEEE. ISBN 978-1-66546-946-3.

Ruan, L.; Chen, B.; Li, J.; and Lam, M.-L. 2021. AIFNet: All-in-Focus Image Restoration Network Using a Light Field-Based Dataset. *IEEE Transactions on Computational Imaging*, 7: 675–688.

Saharia, C.; Ho, J.; Chan, W.; Salimans, T.; Fleet, D. J.; and Norouzi, M. 2022. Image Super-Resolution Via Iterative Refinement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1–14.

Shi, J.; Xu, L.; and Jia, J. 2014. Discriminative Blur Detection Features. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, 2965–2972.

Shi, W.; Caballero, J.; Huszár, F.; Totz, J.; Aitken, A. P.; Bishop, R.; Rueckert, D.; and Wang, Z. 2016. Real-Time Single Image and Video Super-Resolution Using an Efficient Sub-Pixel Convolutional Neural Network. arxiv:1609.05158.

Sohl-Dickstein, J.; Weiss, E. A.; Maheswaranathan, N.; and Ganguli, S. 2015. Deep Unsupervised Learning Using Nonequilibrium Thermodynamics. arxiv:1503.03585.

Son, H.; Lee, J.; Cho, S.; and Lee, S. 2021. Single Image Defocus Deblurring Using Kernel-Sharing Parallel Atrous Convolutions. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2622–2630. Montreal, QC, Canada: IEEE. ISBN 978-1-66542-812-5.

Song, J.; Meng, C.; and Ermon, S. 2022. Denoising Diffusion Implicit Models. arxiv:2010.02502.

Wang, Z.; Bovik, A.; Sheikh, H.; and Simoncelli, E. 2004. Image Quality Assessment: From Error Visibility to Structural Similarity. *IEEE Transactions on Image Processing*, 13(4): 600–612.

Whang, J.; Delbracio, M.; Talebi, H.; Saharia, C.; Dimakis, A. G.; and Milanfar, P. 2022. Deblurring via Stochastic Refinement. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 16272–16282. New Orleans, LA, USA: IEEE. ISBN 978-1-66546-946-3.

Xia, B.; Zhang, Y.; Wang, S.; Wang, Y.; Wu, X.; Tian, Y.; Yang, W.; and Gool, L. V. 2023. DiffIR: Efficient Diffusion Model for Image Restoration. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 13049–13059.

Xie, E.; Wang, W.; Yu, Z.; Anandkumar, A.; Alvarez, J. M.; and Luo, P. 2021. SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers. In *Advances in Neural Information Processing Systems*, volume 34, 12077–12090. Curran Associates, Inc.

Xiong, Y.; Li, Z.; Chen, Y.; Wang, F.; Zhu, X.; Luo, J.; Wang, W.; Lu, T.; Li, H.; Qiao, Y.; Lu, L.; Zhou, J.; and Dai, J. 2024. Efficient Deformable ConvNets: Rethinking Dynamic and Sparse Operator for Vision Applications. arxiv:2401.06197.

Xu, B.; Wang, N.; Chen, T.; and Li, M. 2015. Empirical Evaluation of Rectified Activations in Convolutional Network. arxiv:1505.00853.

Ye, T.; Zhang, Y.; Jiang, M.; Chen, L.; Liu, Y.; Chen, S.; and Chen, E. 2022. Perceiving and Modeling Density for Image Dehazing. In *European Conference on Computer Vision*, 130–145. Springer.

Yuan, L.; Sun, J.; Quan, L.; and Shum, H. 2007. Image deblurring with blurred/noisy image pairs. *ACM SIGGRAPH 2007 papers*.

Zhai, J.; Zeng, P.; Ma, C.; Chen, J.; and Zhao, Y. 2023. Learnable Blur Kernel for Single-Image Defocus Deblurring in the Wild. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(3): 3384–3392.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. arxiv:1801.03924.

Zhuo, S.; and Sim, T. 2011. Defocus Map Estimation from a Single Image. *Pattern Recognition*, 44(9): 1852–1858.