Multi-Level Blur-Aware Stable Diffusion for Region-Adaptive Defocus Deblurring

Xiaopan Li[1*], Yi Jiang[2*], Shiqian Wu[3,4†], Shoulie Xie[5], Sos Agaian[6]
[1]School of Information Engineering, Hubei University of Economics, China
[2]Wuhan Guide Infrared Co., Ltd.
[3]Institute of Advanced Displays and Imaging, Henan Academy of Sciences, China
[4] School of Electronic Information, Wuhan University of Science and Technology, China
[5] Institute for Infocomm Research, A*STAR, Singapore
[6] College of Staten Island, City University of New York, USA
lxp2017@wust.edu.cn, jiangy@guideir.com, shiqian.wu@wust.edu.cn, slxie@i2r.a-star.edu.sg, sos.agaian@csi.cuny.edu

Abstract

Defocus blur, common in shallow depth-of-field photography, varies across image regions and is challenging to accurately estimate and restore. Existing deblurring methods often struggle to capture fine structural textures and do not effectively adapt to regional differences in blur. We propose Multi-Level Blur-Aware Stable Diffusion (MBSD), a novel framework that explicitly integrates regional blur recognition into a diffusion-based image restoration process. MBSD assigns blur-level labels to image patches using a Patch Blur Annotator (PBA), guiding a Multi-Scale Blur Estimator (MSBE) to predict soft blur probabilities and generate routing weights. These weights control a Blur-Adaptive Expert Mixer (BAEM), which adaptively combines features based on local blur severity. The features are then passed to a text-to-image diffusion model via a cross-attention mechanism, enabling region-specific restoration. Extensive experiments on public benchmarks demonstrate that MBSD delivers superior perceptual quality while maintaining competitive PSNR and SSIM, consistently outperforming state-of-the-art methods.

Introduction

Defocus blur occurs when scene points are imaged as circles of confusion (CoC) on the camera sensor, caused by a mismatch between the scene's depth and the camera's depth of field (DoF) (Liang et al. 2024). The size and shape of the blur depend on camera settings, such as aperture design and lens characteristics, and vary significantly across the image depending on the scene geometry (Quan et al. 2024). Defocus deblurring aims to restore image sharpness by removing such blur, with applications spanning photography, computational imaging, and high-level vision tasks (Li et al. 2024).

Traditional defocus deblurring methods often rely on a two-stage process (Liu et al. 2020; Xu, Quan, and Ji 2017): first estimating the spatially varying blur kernel, usually assumed to be Gaussian (Karaali and Jung 2017) or disk-shaped (D'Andrès et al. 2016), followed by non-blind deconvolution (Krishnan and Fergus 2009). However, real-world defocus blur is often more complex than these simple models assume, leading to inaccurate estimations and suboptimal restoration, especially in scenes with mixed blur levels or intricate content (Tang et al. 2024).

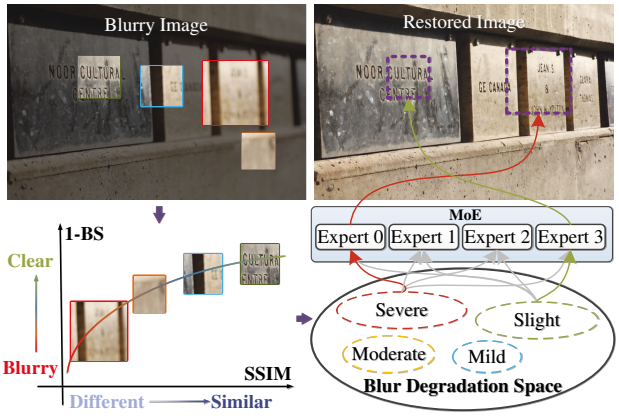


Figure 1: Region-adaptive defocus deblurring via blur-aware expert routing within a Mixture-of-Experts (MoE) framework. To address spatially varying defocus, we model a blur degradation space that encodes local blur severity and informs dynamic expert routing within the MoE. This enables targeted, region-specific restoration, resulting in perceptually coherent and spatially consistent outputs.

Recent learning-based methods (Ruan et al. 2022; Abuolaim and Brown 2020) directly map blurred inputs to sharp outputs using end-to-end deep networks, achieving notable improvements. To better handle spatially varying blur, many works have explored dynamic kernels (Lee et al. 2021; Son et al. 2021), multi-scale representations (Quan, Yao, and Ji 2023; Quan, Wu, and Ji 2023), attention mechanisms (Tang et al. 2024), and Mixture-of-Experts (MoE) architectures (Liang, Zeng, and Zhang 2022). However, most existing models process images uniformly across pixels and lack explicit modeling of local blur severity, which limits their ability to restore severely degraded or low-texture regions (Zhang and Zhai 2022). Patch-based strategies provide a promising alternative by enabling localized blur estimation and region-specific processing. This is especially beneficial

*These authors contributed equally.
†Corresponding authors.

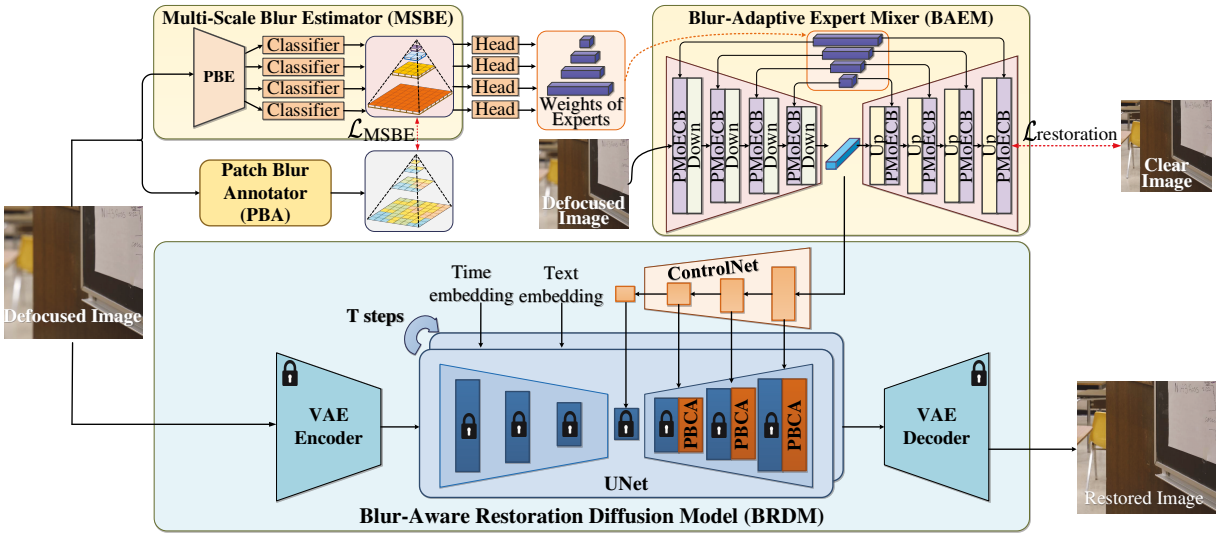


Figure 2: Overview of the proposed MBSD framework. Given a blurry input image, the Patch Blur Annotator (PBA) first generates discrete blur-level labels, which supervise the Multi-Scale Blur Estimator (MSBE) to predict expert weights. These weights guide the Patch-level MoE Convolution Blocks (PMoECB) in the Blur-Adaptive Expert Mixer (BAEM) to extract blur-aware features f_{PBF}. A fidelity restoration loss $\mathcal{L}$restoration is applied to preserve image fidelity in f_{PBF}. Finally, the Blur-Aware Restoration Diffusion Model (BRDM) leverages f_{PBF} as guidance to produce a high-quality restored image.

in defocus deblurring, where blur severity varies with scene depth and object distance (Mao, Li, and Wang 2024). By estimating blur levels at the patch level, models can better prioritize heavily blurred regions.

In parallel, diffusion models such as Stable Diffusion (SD) (Rombach et al. 2022) have shown impressive potential in image restoration tasks like super-resolution (Wu et al. 2024), deblurring (Chen et al. 2025), and dehazing (Yang et al. 2024b), leveraging rich semantic priors to generate perceptually realistic content. However, existing diffusion-based methods are typically trained on clean images and treat degradations globally, without considering spatial variation. This often leads to over-smoothing in relatively sharp regions and hallucinated structures in severely blurred areas (Yang et al. 2024a). These limitations highlight the importance of incorporating explicit blur awareness into the diffusion process to enable region-adaptive restoration.

To address these issues, we propose Multi-Level Blur-Aware Stable Diffusion (MBSD), an end-to-end framework that explicitly integrates regional blur awareness into the diffusion-based restoration process. Specifically, we adopt a re-blurred strategy to assess patch-wise blur severity and assign discrete blur-level labels. These labels supervise multi-scale classifiers to predict soft blur probabilities, capturing the spatial distribution of defocus. Based on these probabilities, adaptive routing weights are computed to activate patch-wise MoE modules, where each expert is specialized for a specific blur level (see Fig. 1). To further enhance feature quality, a fidelity restoration pretext task is introduced to guide blur-aware feature learning toward high-fidelity reconstruction. The resulting features, enriched with both blur and fidelity priors, are injected into the Stable Diffusion U-Net via ControlNet and a cross-attention mechanism, enabling spatially adaptive and perceptually consistent restoration. The entire framework is trained end-to-end using a combination of losses, including cross-entropy, fidelity restoration, and diffusion noise prediction, to jointly improve restoration fidelity and perceptual quality. By explicitly modeling spatial degradation and combining it with the generative power of diffusion models, MBSD achieves high-fidelity and region-aware defocus deblurring.

The main contributions of the proposed method can be summarized as follows:

- A blur-level annotator for creating discrete regional blur labels, guiding multi-scale classifiers to predict soft, patch-wise blur-level probabilities.
- A blur-adaptive expert mixer that dynamically routes features based on local blur severity, improving the model's ability to handle different blur conditions.
- A integration of blur-aware features into Stable Diffusion through ControlNet and cross-attention, enabling spatially adaptive and high-quality image restoration.

Method

The proposed MBSD framework, as illustrated in Fig. 2, consists of four key components: 1) Patch Blur Annotator (PBA), which categories each image patch to one of four discrete blur levels based on computed blur scores; 2) Multi-Scale Blur Estimator (MSBE), which generates multi-scale probabilistic blur maps under the supervision of PBA annotations; 3) Blur-Adaptive Expert Mixer (BAEM), which extracts blur-aware features by dynamically routing information based on the estimated blur priors from MSBE; and 4) Blur-Aware Restoration Diffusion Model (BRDM), which integrates the extracted features into the denoising process of

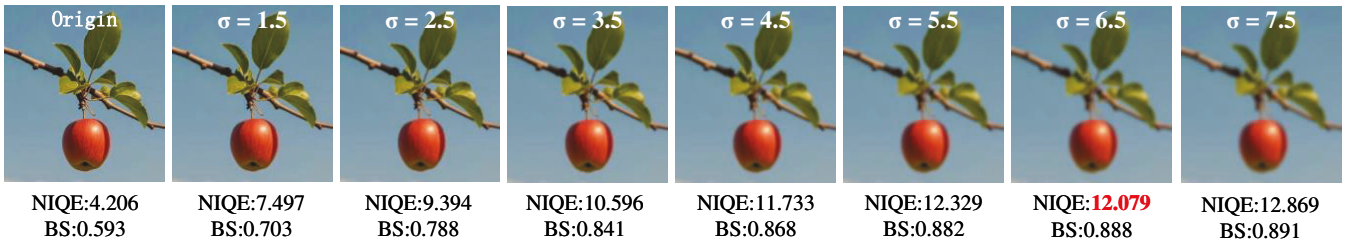


Figure 3: Comparison between the proposed Blur Score (BS) and NIQE (Zhang, Zhang, and Bovik 2015). With increasing Gaussian blur σ, BS rises monotonically, while NIQE fails to do so and gives incorrect scores in the red-marked area.

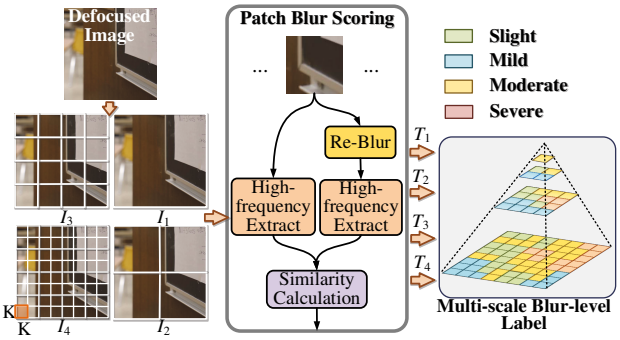


Figure 4: Illustration of the proposed Patch Blur Annotator (PBA). The input image is divided into multi-scale patches, each scored based on the SSIM between its original and re-blurred high-frequency components, and then discretized into four blur levels using predefined thresholds.

a diffusion model for spatially adaptive restoration. By embedding region-specific blur priors into the diffusion framework, MBSD achieves enhanced restoration fidelity and perceptual quality in defocus deblurring.

Patch Blur Annotator (PBA). Accurate blur estimation is critical for spatially varying image restoration. But as shown in Fig. 3, Naturalness Image Quality Evaluator (NIQE) (Zhang, Zhang, and Bovik 2015), a widely used no-reference metrics, often exhibits non-monotonic behavior with increasing blur strength, resulting in unreliable assessments. To address this, we propose a monotonic and interpretable blur scoring strategy, as shown in Fig. 4. Given an image $I \in \mathbb{R}^{H\times W}$, we divide it into $N = HW/K^2$ non-overlapping patches $P^{i,j}$ of size $K\times K$, where $i \in [1, H/K]$ and $j \in [1, W/K]$. The blur score $\mathbb{M}(P)$ is derived by the re-blurred theory: sharp patches lose more high-frequency content when blurred, whereas already blurry patches exhibit less change. Specifically, we first apply a Gaussian filter $K_G(\sigma, r)$ to obtain the re-blurred patch P_r, then extract high-frequency components from both P and P_r using a Laplacian filter K_L. The blur score is defined as:

$$\mathbb{M}(P) = SSIM(K_L * (K_G(\sigma, r) * P), K_L * P), \quad (1)$$

where $*$ denotes convolution, σ and r are the standard deviation and radius of the Gaussian filter, respectively. A higher $\mathbb{M}(P)$ indicates a higher degree of blur.

To incorporate multi-scale context, we generate L scales by varying the patch size. At scale l, patches are sized $2^{l-1}K \times 2^{l-1}K$, and each is assigned a blur score. The resulting blur score map at scale l is:

$$S(I_l) = \begin{bmatrix} \mathbb{M}(P_l^{1,1}) & \cdots & \mathbb{M}(P_l^{1,W_l}) \\ \vdots & \ddots & \vdots \\ \mathbb{M}(P_l^{H_l,1}) & \cdots & \mathbb{M}(P_l^{H_l,W_l}) \end{bmatrix}, \quad (2)$$

where $H_l = \frac{H}{2^{l-1}K}$ and $W_l = \frac{W}{2^{l-1}K}$.

Since estimating precise continuous blur scores for each patch is inherently difficult, we discretize the scores into four distinct levels: slight, mild, moderate, and severe. The discrete blur level label $T(\mathbb{M}(P))$ is defined as:

$$T(\mathbb{M}(P)) = \begin{cases} 0, & \text{if } \mathbb{M}(P) < 0.25, \\ 1, & \text{if } \mathbb{M}(P) < 0.35, \\ 2, & \text{if } \mathbb{M}(P) < 0.75, \\ 3, & \text{otherwise.} \end{cases} \quad (3)$$

The thresholds $\{0.25, 0.35, 0.75\}$ are empirically set to ensure alignment with perceptual blur distinctions.

Multi-Scale Blur Estimator (MSBE). The MSBE utilizes the discrete blur-level labels generated by PBA as supervision to learn probabilistic blur predictions across multiple scales. As illustrated in Fig.2, MSBE consists of a Patch Blur-aware Extractor (PBE) and L scale-specific classifier branches. The PBE is composed of 8 sequential blocks, each consisting of a 3×3 convolution followed by a ReLU activation, which encodes the entire image into a shared deep feature map f_P. This shared feature is then processed by $L = 4$ classifier branches, each corresponding to a specific patch scale. Each classifier branch consists of an average pooling layer (with downsampling rates of $\{1, 2, 4, 8\}$, respectively) followed by two 1×1 convolutions. The channel dimension is finally reduced to 4, corresponding to the four discrete blur levels. The classifier output at each scale is denoted as $R_l \in \mathbb{R}^{4\times H_l\times W_l}$, where $R_{l,e}^{i,j}$ represents the probability that patch $P_l^{i,j}$ belongs to blur level e. The MSBE is trained using a multi-scale cross-entropy loss:

$$\mathcal{L}_{\text{MSBE}} = -\frac{1}{\sum_{l=1}^{L} H_l W_l} \sum_{l=1}^{L} \sum_{i=1}^{H_l} \sum_{j=1}^{W_l} \sum_{e=1}^{4} y_{l,e}^{i,j} \log R_{l,e}^{i,j}, \quad (4)$$

where $y_{l,e}^{i,j}$ is the one-hot encoded ground-truth label $T(\mathbb{M}(P_l^{i,j}))$. This design enables the network to perform patch-wise blur classification across multiple scales.

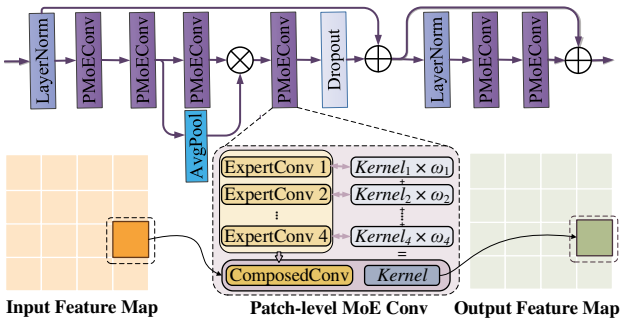


Figure 5: Patch-level MoE Convolution Blocks (PMoECB). The input feature map is divided into fixed-size patches, and each patch is processed by a weighted combination of N experts with shared architecture but independent parameters.

Blur-Adaptive Expert Mixer (BAEM). BAEM adopts a Mixture-of-Experts (MoE) mechanism, where each expert is specialized to handle a specific level of blur severity. To enable adaptive expert weighting, BAEM utilizes the blur probability maps R predicted by the MSBE. Specifically, for four different scales, MSBE produces probability maps $R_l \in \mathbb{R}^{4\times H_l \times W_l}$ for $l = 1, 2, 3, 4$, where each map encodes the probabilities of four discrete blur levels. These maps are individually processed by a lightweight head, composed of two 1×1 convolutional layers, to generate the expert weight map $\omega_l \in \mathbb{R}^{4\times H_l \times W_l}$. This design enables the model to learn spatially adaptive mixing weights for each expert based on the blur-level predictions. To align with multi-scale blur estimation, BAEM incorporates L downsampling/upsampling branches for consistent coarse-to-fine modeling. The backbone of BAEM is built upon NAFNet (Chen et al. 2022), where each layer is augmented with a proposed Patch-level MoE Convolution Block (PMoECB). As illustrated in Fig. 5, PMoECB extends the original NAFBlock by replacing the standard convolution with a dynamically weighted expert convolution, composed of four parallel expert branches with independent parameters:

$$f_o = \left(\sum_{i=1}^{4} \omega_i \cdot Kernel_i\right) * fi, \tag{5}$$

where f_i and f_o are the input and output feature maps, $Kernel_i$ denotes the convolution parameters of the i-th expert, and ω_i is its corresponding adaptive weight.

The encoder of BAEM extracts patch-wise blur-aware features, denoted as f_{PBF}, which are used to guide the subsequent restoration diffusion model. To ensure f_{PBF} retains fidelity information, we introduce a fidelity restoration pretext task: f_{PBF} is decoded into a restored image $\hat{I}$ and supervised using pixel-wise Mean Squared Error (MSE) loss:

$$\mathcal{L}_{\text{restoration}} = \|\hat{I} - I\|_2^2, \tag{6}$$

where I is the ground truth clear image.

Blur-Aware Restoration Diffusion Model (BRDM). The BRDM integrates blur-aware features into the image generation process by leveraging the conditional control capability of ControlNet (Zhang, Rao, and Agrawala 2023). As shown in Fig. 2, we initialize ControlNet with the encoder of a pre-trained Stable Diffusion (SD) model. To obtain reliable semantic prompts, we adopt the DAPE (Wu et al. 2024) to generate textual descriptions. These prompts are encoded by the built-in text encoder of SD to produce text embeddings, which are incorporated into the generation process via the original text cross-attention (TCA) blocks. In addition to text embeddings, patch-wise blur-aware features f_{PBF} extracted by BAEM are introduced into ControlNet as conditional inputs. Intermediate features from multiple ControlNet layers are then injected into the SD U-Net through Patch-wise Blur-aware Cross-Attention (PBCA) modules, positioned after the original TCA blocks.

During training, BRDM follows the standard diffusion process. Given a clear image, its VAE encoder produces the latent representation z_0, which is gradually perturbed with Gaussian noise to obtain z_τ at timestep τ. With noise ϵ_τ and conditional features f_{PBF}, the BRDM denoising network Φ_θ is trained to predict the added noise:

$$\mathcal{L}_{\text{noise}} = \|\epsilon_\tau - \Phi_\theta(z_\tau, \tau, f_{PBE})\|_2^2. \tag{7}$$

Training Losses. All modules, including MSBE, BAEM, and BRDM, are jointly trained in an end-to-end manner. Specifically, $\mathcal{L}_{\text{MSBE}}$ supervises MSBE to predict soft blur-level probabilities, while $\mathcal{L}_{\text{restoration}}$ enhances the guidance quality of BAEM. For BRDM, we adopt the standard noise prediction loss $\mathcal{L}_{\text{noise}}$. The total training loss is defined as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{noise}} + \lambda_1 \mathcal{L}_{\text{MSBE}} + \lambda_2 \mathcal{L}_{\text{restoration}}, \tag{8}$$

where λ_1 and λ_2 are hyperparameters that balance the contribution of each loss term. Both are empirically set to 1 in all experiments.

Experiments

Experimental Setting

Implementation Details. In the PBA module, we adopt four scales ($L = 4$), using patch sizes of 512×512, 256×256, 128×128, and 64×64 at each scale. Re-blurred references are generated using a Gaussian filter with a radius of 11 and a standard deviation of 3.6. For the SD model, we use the publicly available SD-2 base model[1] as the pretrained backbone. Training is performed on 512×512 image patches for 1200k iterations using the Adam optimizer (Kingma 2015). We set the learning rate to 5×10^{-5}, and use a batch size of 1. During inference, we adopt 20 sampling steps and apply the LR embedding strategy proposed in (Wu et al. 2024). All experiments are conducted on a single NVIDIA RTX A6000 GPU with 48 GB of memory.

Datasets. We train our model on the DPDD dataset (Abuolaim and Brown 2020), which contains 500 unique scenes, split into 350 training pairs, 74 validation pairs, and 76 testing pairs. To assess the generalization ability, we evaluate the model on the RealDOF dataset (Lee et al. 2021), which includes 50 scenes captured with a Sony a7R IV camera. All RealDOF images are downscaled to 1120×1680 for computational efficiency. Additionally, we perform perceptual evaluation on the PixelDP (Abuolaim and Brown 2020)

[1]https://huggingface.co/stabilityai/stable-diffusion-2-base

Datasets	Metrics	IFAN	Restormer	INIKNet	NRKNet	DefocusGAN	DEDDNet	P2IKT	ViTDeblur	RDDM	Ours
DPDD	PSNR ↑	25.37	25.98	26.06	26.11	25.52	25.57	26.29	26.11	25.97	**26.55**
	SSIM ↑	0.789	0.811	0.803	0.810	0.791	0.794	0.807	**0.814**	0.811	0.800
	LPIPS ↓	0.216	0.178	0.185	0.210	0.165	0.162	0.191	0.201	0.166	**0.106**
	DISTS ↓	0.1331	0.1204	0.1234	0.1396	0.0965	0.0975	0.1311	0.1360	-	**0.0611**
	MUSIQ ↑	59.39	61.78	59.32	59.47	60.43	60.65	59.50	60.72	-	**61.92**
	MANIQA ↑	0.466	0.487	0.471	0.464	0.485	0.491	0.474	0.482	-	**0.542**
	CLIPIQA ↑	0.403	0.441	0.410	0.406	0.520	0.504	0.422	0.403	-	**0.605**
RealDOF	PSNR ↑	24.71	25.09	25.23	25.15	24.07	24.37	25.78	25.14	25.03	**26.32**
	SSIM ↑	0.748	0.762	0.765	0.768	0.734	0.748	**0.787**	0.769	0.772	0.769
	LPIPS ↓	0.306	0.285	0.287	0.338	0.280	0.264	0.235	0.295	0.271	**0.179**
	DISTS ↓	0.1455	0.1267	0.1316	0.1488	0.1255	0.1205	0.1389	0.1257	-	**0.0826**
	MUSIQ ↑	40.31	**48.30**	41.28	38.74	42.35	43.66	40.40	43.81	-	43.21
	MANIQA ↑	0.365	0.419	0.397	0.365	0.399	0.408	0.381	0.392	-	**0.437**
	CLIPIQA ↑	0.241	0.297	0.278	0.277	0.284	0.266	0.275	0.284	-	**0.318**
PixelDP	MUSIQ ↑	45.07	45.50	43.02	45.74	40.24	40.06	44.91	46.01	-	**46.04**
	MANIQA ↑	0.403	0.432	0.426	0.432	0.439	0.444	0.434	0.429	-	**0.448**
	CLIPIQA ↑	0.312	0.377	0.266	0.320	0.376	0.374	0.370	0.310	-	**0.384**

Table 1: Quantitative comparison of DPDD-trained models on three datasets. The best and second best results of each metric are highlighted in **bold** and underlined, respectively.

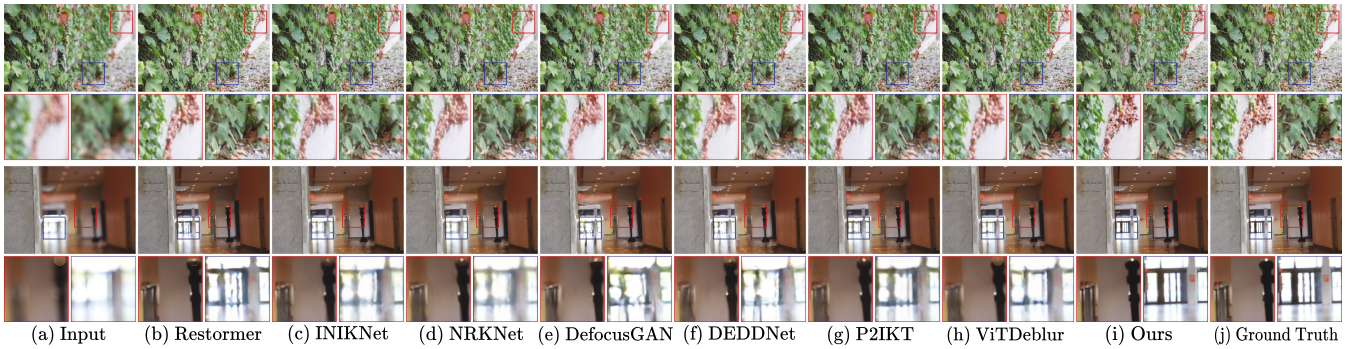


Figure 6: Qualitative results on DPDD (Abuolaim and Brown 2020) dataset among Restormer (Zamir et al. 2022), INIKNet (Quan, Yao, and Ji 2023), NRKNet (Quan, Wu, and Ji 2023), DefocusGAN (Zhai et al. 2023), DEDDNet (Zhai et al. 2024), P2IKT (Tang et al. 2024), ViTDeblur (Liang et al. 2024) and Ours.

dataset, which lacks all-in-focus ground truth.

Evaluation Metrics. We evaluate defocus deblurring performance using both full-reference and no-reference metrics. For fidelity assessment, we use PSNR and SSIM (Wang et al. 2004). For perceptual quality evaluation, we adopt LPIPS (Zhang et al. 2018) and DISTS (Ding et al. 2020). Additionally, we include several no-reference metrics: MUSIQ (Ke et al. 2021), MANIQA (Yang et al. 2022), and CLIPIQA (Wang, Chan, and Loy 2023).

Compared Methods. We compare our MBSD framework with several state-of-the-art defocus deblurring methods, including IFAN (Lee et al. 2021), Restormer (Zamir et al. 2022), INIKNet (Quan, Yao, and Ji 2023), NRKNet (Quan, Wu, and Ji 2023), DefocusGAN (Zhai et al. 2023), DEDDNet (Zhai et al. 2024), P2IKT (Tang et al. 2024), ViTDeblur (Liang et al. 2024), and RDDM (Feng et al. 2025). For fair comparison, we use the publicly released pretrained models provided by the authors. For RDDM (Feng et al. 2025), we report the quantitative results from the original paper due to the absence of released code and models.

Deblurring Performance Comparisons

Quantitative Comparisons. Table 1 reports quantitative results of our MBSD with several state-of-the-art defocus deblurring methods on three benchmark datasets: DPDD, RealDOF, and PixelDP. MBSD consistently delivers competitive or superior performance across all metrics, demonstrating strong generalization. On the DPDD dataset, it achieves the highest PSNR, outperforming P2IKT by +0.26 dB, and yields the lowest LPIPS and DISTS, improving over the second-best by 0.056 and 0.0354, respectively. While ViTDeblur reports the highest SSIM (0.814), MBSD achieves a comparable SSIM (0.800) with notably better perceptual quality. It also ranks first on all no-reference metrics: MUSIQ, MANIQA, and CLIPIQA. On the RealDOF dataset, MBSD again leads in PSNR, LPIPS, and DISTS, with respective improvements of +0.54 dB, -0.056, and -0.0379. Although P2IKT attains the highest SSIM (0.787),

(a) Input (b) Restormer (c) INIKNet (d) NRKNet (e) DefocusGAN (f) DEDDNet (g) P2IKT (h) ViTDeblur (i) Ours (j) Ground Truth

Figure 7: Qualitative results on RealDOF (Lee et al. 2021) dataset among Restormer (Zamir et al. 2022), INIKNet (Quan, Yao, and Ji 2023), NRKNet (Quan, Wu, and Ji 2023), DefocusGAN (Zhai et al. 2023), DEDDNet (Zhai et al. 2024), P2IKT (Tang et al. 2024), ViTDeblur (Liang et al. 2024) and Ours.

(a) Input (b) Restormer (c) INIKNet (d) NRKNet (e) DefocusGAN (f) DEDDNet (g) P2IKT (h) ViTDeblur (i) Ours

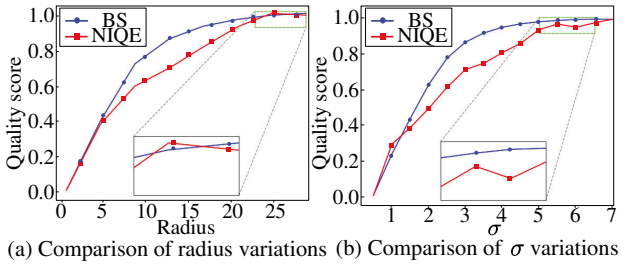
Figure 8: Perceptual quality results on PixelDP (Abuolaim and Brown 2020) dataset among IFANet (Lee et al. 2021), Restormer (Zamir et al. 2022), INIKNet (Quan, Yao, and Ji 2023), NRKNet (Quan, Wu, and Ji 2023), DefocusGAN (Zhai et al. 2023), DEDDNet (Zhai et al. 2024), P2IKT (Tang et al. 2024), ViTDeblur (Liang et al. 2024) and Ours.

BRDM	PBA	MSBE	BAEM	$\mathcal{L}_{\text{restoration}}$	PSNR ↑	SSIM ↑	LPIPS ↓
✓					25.12	0.773	0.143
✓	✓		✓		25.49	0.782	0.123
✓	✓	✓	✓		25.97	0.789	0.109
✓	✓	✓	✓	✓	**26.55**	**0.800**	**0.106**

Table 2: Ablation study on the DPDD dataset to evaluate the effectiveness of each component in our MBSD framework, including the fidelity restoration loss $\mathcal{L}_{\text{restoration}}$.

MBSD shows stronger perceptual fidelity, supported by its top LPIPS, DISTS, MANIQA, and CLIPIQA scores, along with competitive MUSIQ (43.21). For the PixelDP dataset, which lacks full-reference ground truth, MBSD ranks first in all no-reference metrics: MUSIQ (46.04), MANIQA (0.448), and CLIPIQA (0.384), with respective gains of +0.03, +0.004, and +0.007 over the second-best methods.

Qualitative Comparisons. Figs. 6–8 present qualitative comparisons on the DPDD, RealDOF, and PixelDP datasets, respectively. In each figure, two zoom-in regions with varying blur levels are highlighted, where our method consistently delivers visually compelling restorations. In Fig. 6, on the DPDD dataset, MBSD more faithfully restores fine textures and sharp edges than existing methods, particularly in foliage and architectural regions. Competing approaches such as DefocusGAN, DEDDNet, and ViTDeblur often leave residual blur or produce oversmoothed results, especially in regions with complex or transition-level blur. Notably, our MBSD exhibits excellent performance in recovering both severely blurred regions and mildly defocused areas within the same image. In Fig. 7, on the RealDOF dataset, where blur degradation is more realistic and spatially diverse, our method again delivers clear visual improvements. Other methods either leave traces of blur or introduce artifacts such as ringing, halos, or unnatural sharpening. In contrast, our method restores the clean contours and maintains structural alignment, indicating a better understanding of scene geometry and defocus characteristics. Fig. 8 shows results on the PixelDP dataset, which lacks ground truth. Our MBSD produces visually superior outputs with enhanced contrast and local detail, especially in low-light and texture-rich regions. The highlighted zoom-in patches demonstrate that our method maintains consistent structure and text clarity compared to other approaches. These results confirm the effectiveness of our MBSD framework in handling spatially varying blur and generating high-quality deblurred images across diverse defocus scenarios.



(a) Comparison of radius variations (b) Comparison of σ variations

Figure 9: Comparison between our Blur Score (BS) and the NIQE (Zhang, Zhang, and Bovik 2015) metric. (a) plots quality scores with increasing Gaussian blur kernel radius; (b) shows quality scores with increasing Gaussian blur standard deviation (σ). Note that NIQE scores are normalized using min-max scaling.

(a) Input (b) PBA-derived map (c) MSBE-derived map

Figure 10: Visualization comparison between blur-level maps derived from PBA and MSBE. The four blur levels: *slight*, *mild*, *moderate*, and *severe* are represented by blue, green, yellow, and red, respectively.

(a) Input (b) Baseline (c) Ours (d) Ground truth

Figure 11: Visual comparison between our MBSD and the baseline (only BRDM). By leveraging the accurate and adaptive blur representations, our method produces clearer and more visually consistent results.

Ablation Study

Reliability of the Blur Score. To validate the accuracy of the Blur Score (BS), we compare it with the widely used NIQE (Zhang, Zhang, and Bovik 2015) metric under controlled Gaussian blur conditions. As shown in Fig. 9, we apply Gaussian blur with increasing kernel radius (a) and standard deviation σ (b) to 100 clear images from the DPDD dataset. BS exhibits a clear monotonic increase with blur strength, accurately reflecting the blur level. In contrast, NIQE fluctuates non-monotonically due to its reliance on high-level image statistics and sensitivity to scene content. These results demonstrate that BS provides a more stable and interpretable blur measure.

Impact of Blur-Aware Guidance. We designate BRDM as the baseline Controlled Stable Diffusion model, which is guided solely by the original blurry image via ControlNet. To enhance guidance, we introduce discrete blur-level labels from PBA as expert weights in BAEM. As shown in Table 2, this improves PSNR, SSIM, and LPIPS by 0.37, 0.009, and 0.020, respectively. Further, we supervise MSBE using PBA to learn soft blur-level probabilities and feed them into BAEM. Compared to using discrete labels directly, this yields additional improvements of 0.48 dB (PSNR), 0.007 (SSIM), and 0.014 (LPIPS). This is attributed to the ability of MSBE to generate spatially adaptive blur confidence maps, offering a soft and continuous representation of blur severity. As shown in Fig. 10, the first row illustrates that PBA produces hard labels with abrupt transitions, while MSBE captures smoother blur variations that better align with human perception. In the second row, MSBE clearly emphasizes the stronger blur on the right side of the image. Fig. 11 shows qualitative comparisons between the baseline without blur-aware guidance and our method with probabilistic blur-aware guidance. The baseline often produces oversmoothed or distorted results due to limited guidance. In contrast, integrating MSBE with PBA supervision and BAEM allows BRDM to generate more faithful and perceptually consistent restorations.

Effectiveness of Fidelity Restoration Loss. To validate the contribution of the fidelity restoration loss $\mathcal{L}$restoration, we conduct an ablation study by removing it from the training objective. As shown in Table 2, introducing $\mathcal{L}$restoration improves PSNR from 25.97 to 26.55 (+0.58 dB), increases SSIM from 0.789 to 0.800, and reduces LPIPS from 0.118 to 0.106. These results demonstrate that $\mathcal{L}_{\text{restoration}}$ effectively enhances structural fidelity and perceptual quality by providing additional supervision.

Conclusion

In this paper, we present MBSD, a novel framework for addressing spatially varying defocus deblurring by explicitly modeling local blur through both coarse and fine-grained degradation priors. The proposed blur-guided architecture comprises four key components: 1) a PBA that assigns discrete patch-level blur labels to coarsely estimate local blur severity; 2) a MSBE that generates soft blur probability maps at multiple scales to accurately localize blur regions; 3) a BAEM that dynamically routes features based on the estimated blur priors to generate blur-aware representations; and 4) a BRDM that integrates blur-aware guidance into the denoising steps of a diffusion model. Extensive experiments demonstrate that MBSD consistently achieves state-of-the-art performance across both perceptual and distortion-based metrics, confirming its robustness and strong generalization capability in complex real-world defocus scenarios.

Acknowledgments

This work was supported by High-level Talent Research Start-up Project Funding of Henan Academy of Sciences (Project No. 251829061).

References

Abuolaim, A.; and Brown, M. S. 2020. Defocus deblurring using dual-pixel data. In *European Conference on Computer Vision*, 111–126. Springer.

Chen, L.; Chu, X.; Zhang, X.; and Sun, J. 2022. Simple baselines for image restoration. In *European conference on computer vision*, 17–33. Springer.

Chen, Z.; Cui, G.; Zhao, J.; and Nie, J. 2025. CDRM: Controllable diffusion restoration model for realistic image deblurring. *Expert Systems with Applications*, 275: 127009.

Ding, K.; Ma, K.; Wang, S.; and Simoncelli, E. P. 2020. Image quality assessment: Unifying structure and texture similarity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(5): 2567–2581.

D'Andrès, L.; Salvador, J.; Kochale, A.; and Süsstrunk, S. 2016. Non-parametric blur map regression for depth of field extension. *IEEE Transactions on Image Processing*, 25(4): 1660–1673.

Feng, H.; Zhou, H.; Ye, T.; Chen, S.; and Zhu, L. 2025. Residual Diffusion Deblurring Model for Single Image Defocus Deblurring. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 2960–2968.

Karaali, A.; and Jung, C. R. 2017. Edge-based defocus blur estimation with adaptive scale selection. *IEEE Transactions on Image Processing*, 27(3): 1126–1137.

Ke, J.; Wang, Q.; Wang, Y.; Milanfar, P.; and Yang, F. 2021. Musiq: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, 5148–5157.

Kingma, D. P. 2015. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 1–13.

Krishnan, D.; and Fergus, R. 2009. Fast image deconvolution using hyper-Laplacian priors. *Advances in neural information processing systems*, 22.

Lee, J.; Son, H.; Rim, J.; Cho, S.; and Lee, S. 2021. Iterative filter adaptive network for single image defocus deblurring. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2034–2042.

Li, Y.; Xu, R.; Niu, Y.; Guo, W.; and Zhao, T. 2024. Perceptual decoupling with heterogeneous auxiliary tasks for joint low-light image enhancement and deblurring. *IEEE Transactions on Multimedia*, 26: 6663–6675.

Liang, J.; Zeng, H.; and Zhang, L. 2022. Efficient and degradation-adaptive network for real-world image super-resolution. In *European Conference on Computer Vision*, 574–591. Springer.

Liang, P.; Jiang, J.; Liu, X.; and Ma, J. 2024. Decoupling Image Deblurring into Twofold: A Hierarchical Model for Defocus Deblurring. *IEEE Transactions on Computational Imaging*.

Liu, S.; Liao, Q.; Xue, J.-H.; and Zhou, F. 2020. Defocus map estimation from a single image using improved likelihood feature and edge-based basis. *Pattern Recognition*, 107: 107485.

Mao, X.; Li, Q.; and Wang, Y. 2024. Adarevd: Adaptive patch exiting reversible decoder pushes the limit of image deblurring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 25681–25690.

Quan, Y.; Wu, Z.; and Ji, H. 2023. Neumann network with recursive kernels for single image defocus deblurring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5754–5763.

Quan, Y.; Wu, Z.; Xu, R.; and Ji, H. 2024. Deep single image defocus deblurring via gaussian kernel mixture learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Quan, Y.; Yao, X.; and Ji, H. 2023. Single image defocus deblurring via implicit neural inverse kernels. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 12600–12610.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

Ruan, L.; Chen, B.; Li, J.; and Lam, M. 2022. Learning to deblur using light field generated and real defocus images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16304–16313.

Son, H.; Lee, J.; Cho, S.; and Lee, S. 2021. Single image defocus deblurring using kernel-sharing parallel atrous convolutions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2642–2650.

Tang, P.; Xu, Z.; Zhou, C.; Wei, P.; Han, P.; Cao, X.; and Lasser, T. 2024. Prior and prediction inverse kernel transformer for single image defocus deblurring. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 5145–5153.

Wang, J.; Chan, K. C.; and Loy, C. C. 2023. Exploring clip for assessing the look and feel of images. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, 2555–2563.

Wang, Z.; Bovik, A. C.; Sheikh, H. R.; and Simoncelli, E. P. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4): 600–612.

Wu, R.; Yang, T.; Sun, L.; Zhang, Z.; Li, S.; and Zhang, L. 2024. Seesr: Towards semantics-aware real-world image super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 25456–25467.

Xu, G.; Quan, Y.; and Ji, H. 2017. Estimating defocus blur via rank of local patches. In *Proceedings of the IEEE international conference on computer vision*, 5371–5379.

Yang, S.; Wu, T.; Shi, S.; Lao, S.; Gong, Y.; Cao, M.; Wang, J.; and Yang, Y. 2022. Maniqa: Multi-dimension attention

network for no-reference image quality assessment. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 1191–1200.

Yang, T.; Wu, R.; Ren, P.; Xie, X.; and Zhang, L. 2024a. Pixel-aware stable diffusion for realistic image super-resolution and personalized stylization. In *European Conference on Computer Vision*, 74–91. Springer.

Yang, Z.; Yu, H.; Li, B.; Zhang, J.; Huang, J.; and Zhao, F. 2024b. Unleashing the Potential of the Semantic Latent Space in Diffusion Models for Image Dehazing. In *European Conference on Computer Vision*, 371–389. Springer.

Zamir, S. W.; Arora, A.; Khan, S.; Hayat, M.; Khan, F. S.; and Yang, M.-H. 2022. Restormer: Efficient transformer for high-resolution image restoration. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5728–5739.

Zhai, J.; Liu, Y.; Zeng, P.; Ma, C.; Wang, X.; and Zhao, Y. 2024. Efficient Fusion of Depth Information for Defocus Deblurring. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2640–2644. IEEE.

Zhai, J.; Zeng, P.; Ma, C.; Chen, J.; and Zhao, Y. 2023. Learnable blur kernel for single-image defocus deblurring in the wild. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 3384–3392.

Zhang, J.; and Zhai, W. 2022. Blind attention geometric restraint neural network for single image dynamic/defocus deblurring. *IEEE Transactions on Neural Networks and Learning Systems*, 34(11): 8404–8417.

Zhang, L.; Rao, A.; and Agrawala, M. 2023. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, 3836–3847.

Zhang, L.; Zhang, L.; and Bovik, A. C. 2015. A feature-enriched completely blind image quality evaluator. *IEEE Transactions on Image Processing*, 24(8): 2579–2591.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 586–595.