Multi-Domain Multi-Scale Diffusion Model for Low-Light Image Enhancement

Kai Shang[1,2], **Mingwen Shao**[1*], **Chao Wang**[3], **Yuanshuo Cheng**[1], **Shuigen Wang**[4]

[1]School of Computer Science and Technology, China University of Petroleum (East China), China
[2]Shandong Institute of Petroleum and Chemical Technology, China
[3]ReLER, AAII, University of Technology Sydney, Australia
[4]Yantai IRay Technologies Lt. Co., China
skkkyup@163.com, smw278@126.com, chao.wang-11@student.uts.edu.au, cys1294414023@gmail.com, shuigen.wang@iraytek.com

Abstract

Diffusion models have achieved remarkable progress in low-light image enhancement. However, there remain two practical limitations: (1) existing methods mainly focus on the spatial domain for the diffusion process, while neglecting the essential features in the frequency domain; (2) conventional patch-based sampling strategy inevitably leads to severe checkerboard artifacts due to the uneven overlapping. To address these limitations in one go, we propose a Multi-Domain Multi-Scale (MDMS) diffusion model for low-light image enhancement. In particular, we introduce a spatial-frequency fusion module to seamlessly integrates spatial and frequency information. By leveraging the Multi-Domain Learning (MDL) paradigm, our proposed model is endowed with the capability to adaptively facilitate noise distribution learning, thereby enhancing the quality of the generated images. Meanwhile, we propose a Multi-Scale Sampling (MSS) strategy that follows a divide-ensemble manner by merging the restored patches under different resolutions. Such a multi-scale learning paradigm explicitly derives patch information from different granularities, thus leading to smoother boundaries. Furthermore, we empirically adopt the Bright Channel Prior (BCP) which indicates natural statistical regularity as an additional restoration guidance. Experimental results on LOL and LOLv2 datasets demonstrate that our method achieves state-of-the-art performance for the low-light image enhancement task. Codes are available at https://github.com/Oliiveralien/MDMS.

Introduction

Images captured in low-light conditions often suffer additional derivative scenarios such as low contrast and high noise levels, which may not only affect the visual appearance but also the performance on downstream vision tasks (*e.g.*, classification (Dhananjaya, Kumar, and Yogamani 2021), detection (Wang et al. 2022b) and segmentation (Wang et al. 2022a)). Various attempts have been made to improve low-light images and transform them into high-quality images with normal light. Traditional methods such as histogram equalization-based methods (Kaur, Kaur, and Kaur 2011) and Retinex-based methods (Jobson, Rahman, and Woodell 1997; Rahman, Jobson, and Woodell 2004), usually struggle to handle intricate real-world scenarios. Over the last decade, Deep Learning (DL) (LeCun, Bengio, and Hinton 2015) methods have received a surge of recent interest, resulting in several key advances in low-light image enhancement. Although DL-based methods have achieved visually plausible restoration results, these methods still require integrating elaborately crafted prior information or sophisticated network design and training. Recently, diffusion models have gained significant attention, as they learn data distributions by simulating a fixed forward process. These models have demonstrated remarkable performance in various tasks, such as image restoration (Özdenizci and Legenstein 2023; Luo et al. 2023) and low-light Image enhancement (LLIE) (Zhou, Yang, and Yang 2023). However, existing diffusion-based LLIE methods still remain two problems: **On the one hand**, these methods tend to simply model the noise distribution in the spatial domain while ignoring the frequency domain features. As illustrated in Figure 1, we visualize the frequency statistics and the sampling re-

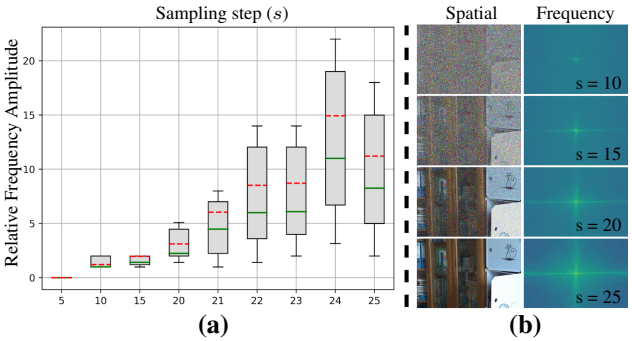

Figure 1: Motivations of our method. (a): Visualizations of the frequency statistics for each sampling step s (Total sampling step $S = 25$). As shown in the box plot, the horizontal axis is the sampling step s, while the vertical axis is the relative frequency amplitude. The red dashed line represents the average frequency value, which gradually increases as the number of steps goes. (b): Visualizations of the spatial (left) and frequency (right) results during sampling.

*Corresponding author: Mingwen Shao

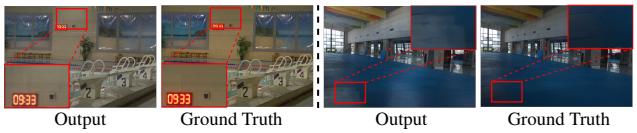

Figure 2: Checkerboard artifacts in previous single-scale sampling methods.

sults for several steps of an existing diffusion-based LLIE method (Özdenizci and Legenstein 2023). It can be observed that the frequency domain features are also progressively optimized during the sampling process (generally following a low-frequency to high-frequency restoration manner), which intuitively motivates us to utilize frequency representations as extra guidance. **On the other hand**, existing methods commonly employ single-resolution patches for both the training and sampling process. Yet this strategy may lead to severe checkerboard artifacts, as shown in Figure 2.

Aiming at these aforementioned problems, we propose a Multi-Domain Multi-Scale Diffusion model, dubbed MDMS-Diffusion. First, to obtain more precise and complete feature distribution, we innovatively fuse the frequency domain information with the spatial domain information to jointly guide the diffusion process. Specifically, we use a multi-domain network to explicitly process the frequency domain information transformed by Fast Fourier Transform (FFT). Notably, though some restoration works (Jiang et al. 2023; Phung, Dao, and Tran 2023) have introduced wavelet transforms (Graps 1995) as a pre-processing operation, it should be noted that our work is the first attempt to integrate frequency representation learning into the diffusion model. Secondly, to subdue the checkerboard artifacts brought by the uneven overlapping problem in single-resolution patches, we introduce a multi-scale sampling strategy. Specifically, we incorporate multi-scale patches to elegantly embellish the boundaries. Correspondingly, we also apply this strategy to enable multi-scale context training and efficient sampling. Additionally, we further customize the Bright Channel Prior to guide the generation process with color and brightness information. Our main contributions are summarized as follows:

- A multi-domain diffusion model is proposed to harvest contextual information from both spatial and frequency space. Such a multi-domain learning paradigm explicitly establishes interactions from the spatial to the frequency domain, thus leading to more precise and faithful restoration for both synthetic and real-world images.
- A multi-scale sampling strategy is introduced to remedy the checkerboard pattern due to the previous single-resolution strategy, which smooths the boundaries by merging restored patches of different resolutions.
- Extensive experiments on LOL and LOLv2 datasets indicate that our method performs favorably against existing low-light image enhancement counterparts.

Related Works

Low-light image enhancement. Low-light image enhancement remains a classical yet challenging low-vision task for many real-world applications. In recent years, generative models (*e.g.*, generative adversarial networks (GAN) (Wang et al. 2023a), normalizing flow (NF) (Wang et al. 2022c) and variational autoencoders (VAE) (Eriksson 2020) have been wildly explored for the low-light enhancement task. However, these DL-based methods either suffer from unstable training (Yang et al. 2022) or lack sufficient representational capacity(Dhariwal and Nichol 2021). Recently, denoising diffusion models (Ho, Jain, and Abbeel 2020; Song, Meng, and Ermon 2020; Rombach et al. 2021) have drawn great attention due to their remarkable performance in various vision tasks. For example, Ozdenizci et al. (Özdenizci and Legenstein 2023) randomly crop patches for training and then reassemble the overlapping patches for general adverse weather restoration. Zhou et al. recently introduce PyDiff (Zhou, Yang, and Yang 2023) with the pyramidal structure and global corrector for low-light image enhancement. However, these approaches primarily learn noise distribution within the spatial domain while disregarding the frequency. Meanwhile, the single-scale patch sampling pattern usually results in inferior artifacts due to the uneven overlap.

Frequency domain analysis. In recent years, several studies (Shao et al. 2023; Phung, Dao, and Tran 2023) have attempted to combine frequency operations with deep learning. For example, Liu et al. (Liu et al. 2020) first apply the Haar wavelet transform to convert the image into the frequency domain for image Demoiréing. Zhou et al. (Zhou et al. 2022) further explore the relationship between the spatial and Fourier domains and introduce Deep Fourier Up-Sampling to extract more global information. However, these frequency-based methods either consider the frequency transform as a preprocessing step or overlook its correlation with the spatial domain. In contrast, our proposed approach seamlessly merges spatial characteristics with frequency information. Combined with the multi-scale sampling strategy, MDMS can deliver more accurate and authentic restoration results.

Methods

According to Figure 1, our proposed method needs to follow a specific learning pattern, which recovers low-frequency information first and subsequently restores more high-frequency information. Existing diffusion models mostly employ spatial domain information for both the forward and backward processes, yet overlook the frequency characteristics. In this paper, we propose a Multi-Domain Multi-Scale (MDMS) diffusion model for low-light image enhancement, as illustrated in Figure 3. The proposed method effectively learns the mapping between low-light and normal-light images by fusing the multi-domain information. In addition, we present a multi-scale strategy to enhance the perceptual capability. Moreover, we integrate the bright channel prior information to effectively guide the generation process.

Preliminary

For the proposed diffusion model, the low-light images y and our proposed Bright Channel Prior (BCP) y_p are taken as conditional inputs to guide the sampling process restoring

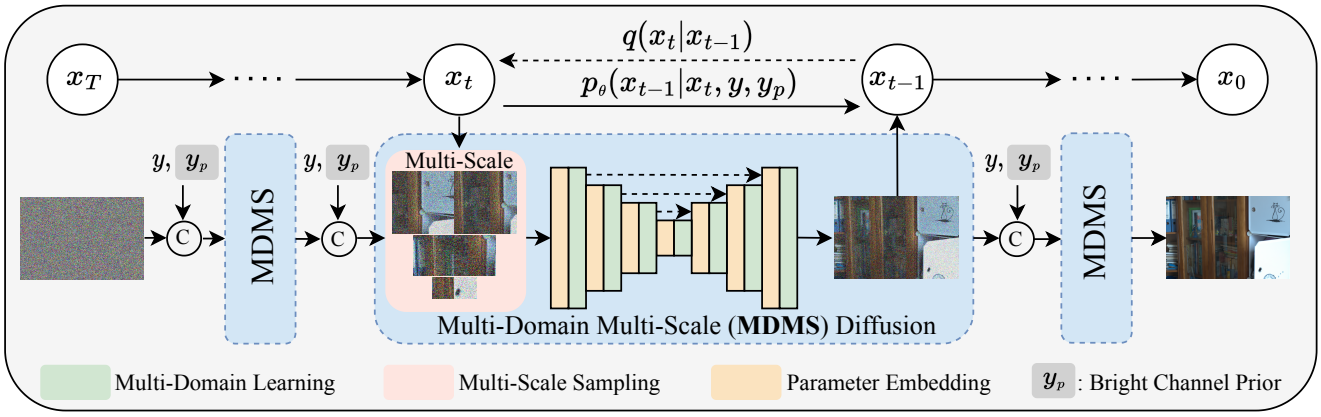

Figure 3: Overview of the forward diffusion and reverse denoising processes for our Multi-Domain Multi-Scale (MDMS) Diffusion framework. Our MDMS model consists of a Multi-Scale Sampling (MSS) and a Multi-Domain Learning (MDL) module. MDMS learns to effectively perform low-light image enhancement conditioned on the illumination prior y_p and the input degraded image y, which first adopts a Bright Channel Prior (BCP) for the auxiliary illumination prior and then restores the multi-scale patches using information from both spatial and frequency domains.

the corresponding normal-light image. The forward process transforms the data distribution of natural image x_0 with a fixed schedule $\{\beta_1, \beta_2, \cdots \beta_T\}$ into a standard Gaussian distribution $x_T \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$:

$$q(\mathbf{x}_{1:T}|\mathbf{x}_0) = \prod_{t=1}^{T} q(\mathbf{x}_t|\mathbf{x}_{t-1}), \tag{1}$$

$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I}), \tag{2}$$

where x_t indicates the corrupted data distribution and β_t represents the pre-defined variance at t step.

The reverse process utilizes a Markov chain to progressively denoise the randomly sampled Gaussian distribution $x_T \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$ and obtain new sampling that conforms to the data distribution x_0:

$$p_\theta(\mathbf{x}_{0:T}|\mathbf{y}, \mathbf{y_p}) = p(\mathbf{x}_T) \prod_{t=1}^{T} p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{y}, \mathbf{y_p}), \tag{3}$$

$$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{y}, \mathbf{y_p}) = \mathcal{N}(\mathbf{x}_{t-1}; \mu_\theta(\mathbf{x}_t, \mathbf{y}, \mathbf{y_p}, t), \sigma_t^2 I), \tag{4}$$

where θ denotes the learnable parameters, while μ_θ and σ_t^2 presents the mean and variance. $p_\theta(\mathbf{x}_t)$ represents the distribution of x_t.

Following DDIM (Song, Meng, and Ermon 2020), we simplify the sampling process by skipping the coefficient of variation. By minimizing the Negative Log-Likelihood (NLL) and re-parameterizing, the general objective function of our MDMS can be formulated as:

$$L_{diff} = \mathbb{E}_{\mathbf{x}_0,t,\epsilon\sim\mathcal{N}(0,\mathbf{I})} \parallel \epsilon - \epsilon_\theta(\mathbf{x}_t, \mathbf{y}, \mathbf{y_p}, t) \parallel^2, \tag{5}$$

where ϵ_θ denotes the neural network, which predicts the corresponding Gaussian noise ϵ based on the variable x_t, the low-light image y, the prior image y_p and timestep t.

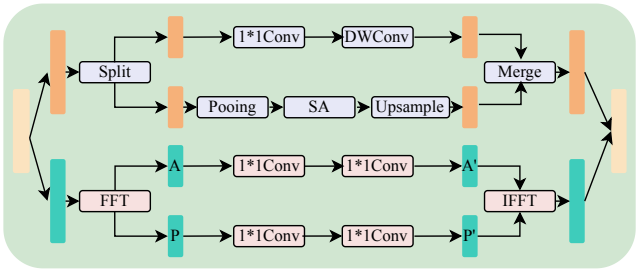

Figure 4: Illustrations of our Multi-Domain Learning (MDL) module, which contains two parallel branches focusing on spatial (top) and frequency (bottom) feature learning, respectively. This module effectively learns more comprehensive representations among multi-domain feature space.

Multi-Domain Learning

As shown in Figure 3, the proposed diffusion model adopts a U-shaped network as the backbone, which aims to predict the noise of each step x_t based on the previous output x_{t+1}, the low-light image y, and the prior information y_p. Unlike existing works, we further delicately curate the U-Net with a novel Multi-Domain Learning (MDL) module and a Parameter Embedding (PE) module to fully exploit the complementary features from multi-domain spaces. The PE module aims at scaling the features and simultaneously embedding the temporal step t with patch information (*e.g.*, patch size and the relative position (Wang et al. 2023c)), while the MDL module is utilized to capture more frequency-sensitive features such as color and textures.

As illustrated in Figure 4, the MDL module consists of two branches: the spatial-domain branch and the frequency branch. Specifically for the frequency-domain branch, we employ the 2D Fast Fourier Transform (FFT) to transform the spatial information into the frequency domain, which can

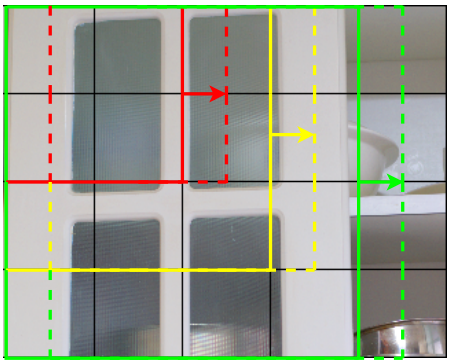

Figure 5: Illustrations of our Multi-Scale Sampling (MSS). The solid lines in red, yellow, and green represent the multi-scale sampling patches, while the dashed lines depict the positions of the sampling patches in the next step.

be formulated as:

$$X(u,v) = \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} x(m,n) \cdot e^{-j2\pi\left(\frac{um}{M}+\frac{vn}{N}\right)}, \quad (6)$$

where $X(u,v)$ and $x(m,n)$ respectively represent the frequency value at (u,v) and the spatial value at (m,n).

The frequency-domain information comprises two components: amplitude and phase. The amplitude component A is primarily related to the intensity, while the phase component P is mainly associated with the details as follows:

$$A : |X(u,v)| = \sqrt{R(u,v)^2 + I(u,v)^2}, \quad (7)$$

$$P : \angle X(u,v) = \arctan\left(\frac{I(u,v)}{R(u,v)}\right), \quad (8)$$

where $R(u,v)$ and $I(u,v)$ represent the real and imaginary parts of $X(u,v)$. To preserve the original structural information, we adopt separate 1×1 convolutions (Lin, Chen, and Yan 2013) on the amplitude and phase components. Subsequently, the features are transformed back to the spatial domain using Inverse Fast Fourier Transform (IFFT). By utilizing frequency domain representation, the network can acquire more extensive features. In the spatial domain branch, we divide the feature channel into halves to obtain both local and global information. We apply depth-separable convolution to extract local features and employ self-attention for the global information. To reduce computational complexity, we perform average pooling on intermediate features to decrease their size, which is then restored to the original size.

Multi-Scale Sampling

Existing patch-based methods (Özdenizci and Legenstein 2023) typically perform the forward and reverse process at a single fixed resolution, aiming at training efficiency and the capability of handling images with arbitrary resolutions. However, this strategy tends to result in checkerboard artifacts as illustrated in Figure 2. This is mainly due to the uneven overlapping problem especially when the patch size cannot be divided by the sampling stride, which usually occurs in traditional Convolutional Neural Networks (CNNs).

Algorithm 1: MDMS Diffusion Model Training

Input: Low-light image y, prior image y_p, normal-light image x_0.

while not converged **do**
  Generate a random size binary mask M_i.
  Crop $x_0^i = M_i \circ x_0$, $y^i = M_i \circ y$, $y_p^i = M_i \circ y_p$.
  Resize $x_0^i = x_0^i \downarrow$, $y^i = y^i \downarrow$, $y_p^i = y_p^i \downarrow$, where $\downarrow$ means downsampling to 64×64.
  Sample $t \sim$ Uniform $\{1, \cdots, T\}$.
  Sample $\epsilon_t \sim \mathcal{N}(0, I)$.
  $x_t^i = \sqrt{\bar{\alpha}}(x_0^i \downarrow) + \sqrt{1-\bar{\alpha}}\epsilon$.
  Take gradient descent step on $\bigtriangledown_\theta \parallel \epsilon - \epsilon_\theta(\mathbf{x}_t^i, \mathbf{y}^i, \mathbf{y_p}^i, t) \parallel^2$.
end while
return θ

Algorithm 2: MDMS Diffusion Model Sampling

Input: Low-light image y, prior image y_p, conditional diffusion model $\epsilon_\theta(\mathbf{x}_t, \mathbf{y}, \mathbf{y_p}, t)$, sampling steps S, patch locations D.

Sample $X_t \sim \mathcal{N}(0, I)$.
for $i = S, \cdots, 1$ **do**
  $t = (i-1) \cdot T/S + 1$.
  $t' = (i-2) \cdot T/S + 1$.
  $\Phi_t = 0, W = 0$.
  for $ps = 64 \times 64, 96 \times 96, 128 \times 128$ **do**
    for $d = 1, \cdots, D$ **do**
      $x_t^d = Crop_{ps}(M_d \circ X_t)$, $y^d = Crop_{ps}(M_d \circ y)$, and $y_p^d = Crop_{ps}(M_d \circ y_p)$.
      $\Phi_{ps} = \Phi_{ps} + M_d \cdot \epsilon_\theta(\mathbf{x}_t^d, \mathbf{y}^d, \mathbf{y_p}^d, t)$.
      $W = W + M_d$.
    end for
    $\Phi_{ps} = \Phi_{ps} \oslash W$, $\oslash$ means element-wise divide.
    $\Phi_t = (\Phi_t + \Phi_{ps})$.
  end for
  $\Phi_t = \Phi_t / 3$.
  $\mathbf{X}_t \leftarrow \sqrt{\bar{\alpha}_{t'}}\left(\frac{\mathbf{X}_t - \sqrt{1-\bar{\alpha}_t} \cdot \mathbf{\Phi}_t}{\sqrt{\bar{\alpha}_t}}\right) + \sqrt{1-\bar{\alpha}_{t'}} \cdot \mathbf{\Phi}_t$.
end for
return X_t

To address this issue, we propose a Multi-Scale Sampling strategy to smooth the boundaries. As shown in Figure 5, we employ three different patch sizes: 64×64, 96×96, 128×128 during the sampling process. At each sampling step t, the intermediate variance x_t is split into overlapped multi-scale patches, with each patch being denoised separately. Subsequently, patches with the same size are merged to match the original image dimensions, and the images synthesized from different scales are eventually fused to obtain x_{t-1}. In the training process, we select image patches of various sizes and uniformly resize them to 64x64 dimensions. Through this random self-sampling pattern, our model is able to learn consistent details from multi-scale patches, each with different receptive fields, without incurring additional computational overhead.

Figure 6: Comparisons of common priors for low-light image enhancement. HE represents the Histogram Equalization. Gamma denotes the Gamma correction. Bright Channel Prior (BCP) clearly provides more natural and faithful lighting information. Please zoom in to see the details.

Bright Channel Prior

Apart from the MDL module for high-frequency feature learning, we also introduce an additional prior, named Bright Channel Prior (BCP), based on the statistics of natural images. Different from the classical dark channel prior (He, Sun, and Tang 2010) for image dehazing, BCP theory (Yan et al. 2017) postulates that natural images contain at least one channel with relatively higher pixel values (*i.e.*, color), which can be written as follow:

$$B(I)(x) = \max_{y \in \Omega(x)} (\max_{c \in (r,g,b)} I^c(y)), \tag{9}$$

where x represents the pixel position, and $\Omega(x)$ represents the region centered at x. I^c denotes the color channel. The bright channel prior is commonly employed as the illumination map for Retinex-based applications (Guo 2016). Motivated by this idea, we intuitively introduce this prior as auxiliary information to guide the diffusion process. We further curate BCP and design the prior for the diffusion model as:

$$y_{\mathrm{p}}(x) = \frac{y(x)}{(\max_{c \in (r,g,b)} I^c(x) + \epsilon)}, \tag{10}$$

where ϵ is a constant introduced to prevent division by zero. In this paper, ϵ is empirically set as 0.1. As shown in Figure 6, our BCP preserves more color and texture details through simple preprocessing, compared with other common-used priors.

Training and Sampling

Training. Algorithm 1 shows the specific training procedure of the proposed MDMS. The Multi-Domain Learning (MDL) module aims to capture subtle frequency-aware features that cannot be noticed in the space domain. The Bright Channel Prior (BCP) aims to inject more natural priors (*e.g.*, color and illumination maps) into the degraded input.

Sampling. Algorithm 2 shows the specific sampling procedure of the proposed MDMS. We adopt an accelerated deterministic sampling approach DDIM (Song, Meng, and Ermon 2020) to reduce the number of sampling timesteps. The Multi-Scale Sampling (MSS) strategy is proposed to mitigate boundary artifacts between adjacent patches.

Training Loss. As outlined in Eq. (5) and Algorithm 1, we employ a straightforward L_2 loss on the predicted noise map to train the denoising network.

Methods	PSNR ↑	SSIM ↑	LPIPS ↓
Zero-DCE (Guo et al. 2020)	14.86	0.562	0.335
DRBN (Yang et al. 2020)	15.15	0.492	0.339
RUAS (Liu et al. 2021)	16.40	0.503	0.270
RetinexNet (Wei et al. 2018)	16.77	0.425	0.474
TBEFN (Lu and Zhang 2020)	17.35	0.777	0.210
EnlightenGAN (Jiang et al. 2021)	17.48	0.652	0.322
MBLLEN (Lv et al. 2018)	17.90	0.701	0.234
SGM-Net (Yang et al. 2021)	17.92	0.753	0.296
GLADNet (Wang et al. 2018)	19.72	0.682	0.321
KinD++ (Zhang et al. 2021)	21.80	0.829	0.158
DLN (Wang et al. 2020)	21.94	0.846	0.142
IAT (Cui et al. 2022)	23.38	0.806	0.216
LLFormer (Wang et al. 2023b)	23.65	0.816	0.169
SNR (Xu et al. 2022)	24.61	0.840	0.151
LLFlow (Wang et al. 2022c)	25.01	0.870	0.117
Pydiff (Zhou, Yang, and Yang 2023)	<u>27.07</u>	<u>0.880</u>	<u>0.100</u>
MDMS(ours)	**27.12**	**0.882**	**0.078**

Table 1: Quantitative results on the LOL dataset in terms of PSNR, SSIM and LPIPS. ↑ means higher is better, while ↓ means lower is better. The best performance is marked in bold with the second performance underlined.

Experiments

Experimental Settings

Dataset. The proposed diffusion model is trained on the LOL dataset (Wei et al. 2018), and evaluated on both LOL and LOLv2 dataset (Yang et al. 2021). The LOL dataset contains 500 paired images, with 485 for training and 15 for testing. LOLv2 dataset consists of two subsets: LOLv2-Real and LOLv2-Syn. LOLv2-Real comprises images captured from real-world scenes, including 689 images for training and 100 images for testing. LOLv2-Syn is a synthesized subset obtained by adjusting the Y-channel of RAW images to match low-light distributions, which consists of 900 images for training and 100 images for testing.

Schedules. For our diffusion model, the time-step T is set to 1,000 for the training stage and the implicit sampling step S is set to 25. Furthermore, our model achieves promising results using alternative step (S = 20, 10, 5, 4). For the noise schedule, α is linearly decreased from 0.999 to 0.98.

Training details. We conduct training using 64×64 patches. To correspond with multi-scale sampling and enhance training diversity, we randomly crop patches of size 256×256, 128×128, and 64×64, and resized them to 64×64. We use the Adam optimizer with an initial learning rate of $2e^{-5}$. In addition to the time step t and patch size, we also add parameters including the top-left and bottom-right coordinates of the cropped patches during the parameter embedding to guide the training and sampling process.

Metrics. We adopt three metrics for evaluation: Peak Signal-to-Noise Ratio (PSNR) (Wang et al. 2004), Structural Similarity (SSIM) (Wang et al. 2004) and Learned Perceptual Image Patch Similarity (LPIPS) (Zhang et al. 2018). PSNR is employed to analyze pixel-wise differences, SSIM is used to evaluate structural information similarity, and LPIPS is utilized to assess perceptual consistency.

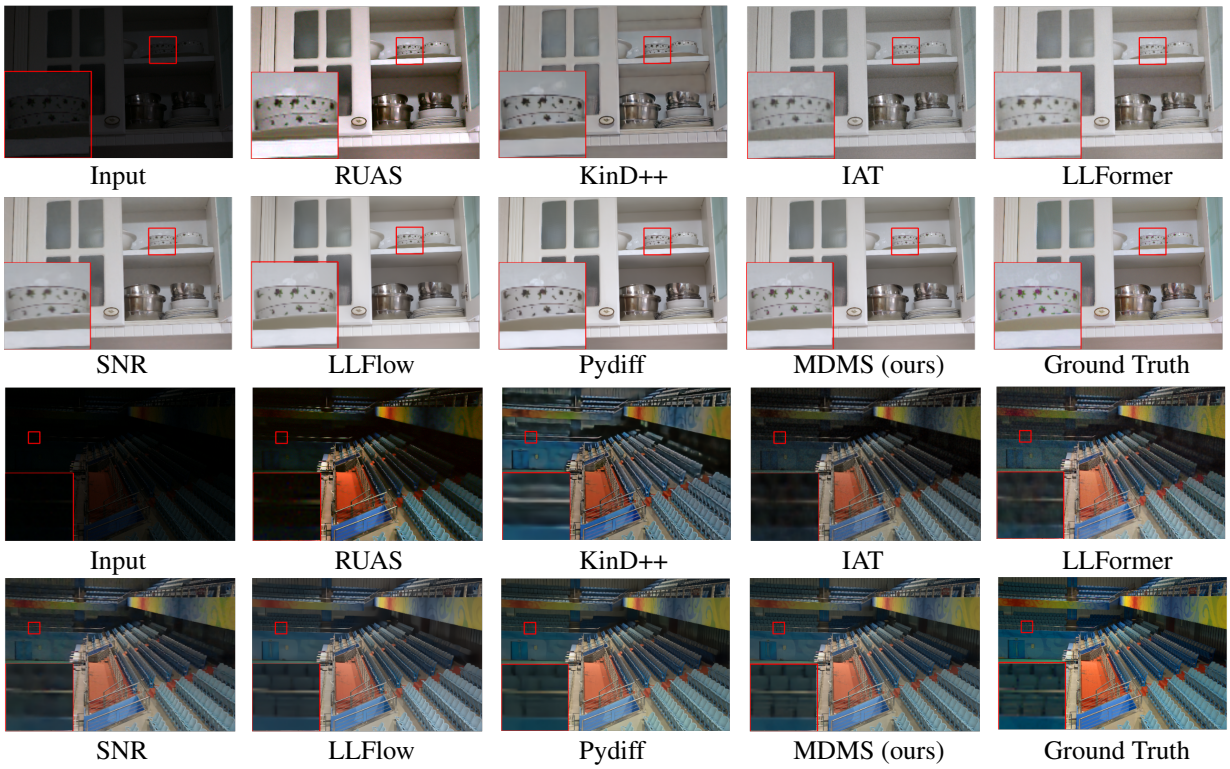

Figure 7: Qualitative comparisons with existing methods on LOL dataset. Please zoom in to see the details.

Comparisons with Existing Methods

LOL Dataset. We first conduct quantitative comparisons of state-of-the-art (SOTA) methods on the LOL dataset, as shown in Table 1. In contrast to the previous SOTA methods, our MDMS approach consistently outperforms them in terms of all evaluated metrics. It is worth noting that our method exhibits a significant improvement in LPIPS, indicating that our approach yields superior visual results. The qualitative comparison results are shown in Figure 7. It can be observed that our method is capable of not only enhancing the overall brightness of the image but also accurately restoring detailed information, such as the colored decoration on the bowl and the boundary of the seat. This can be mainly contributed to the learned characteristics from our MDL module, which facilitates the restoration quality, especially those high-frequency features such as the color information and edge textures. Besides, the utilization of BCP also provides more faithful color guidance (as shown in Figure 6), compared to other priors that simply average the color space such as Histogram Equalization in Pydiff (Zhou, Yang, and Yang 2023) or Gamma correction in KinD++ (Zhang et al. 2021).

LOLv2 Dataset. We also conduct evaluations on the LOLv2 dataset, as shown in Table 2 and Table 3. For a fair comparison, all listed methods are trained on the LOL dataset and evaluated on the LOLv2 dataset. It can be clearly observed that our method achieves the best performance on both real-world and synthetic datasets. Notably, our method outperforms the previous best approach Pydiff (Zhou, Yang, and Yang 2023) on real-world images by a large margin at 2.19dB in PSNR. This is mainly because our additional frequency branch can capture more data-irrelevant features such as intrinsic lightness and structure information for better generalization and adaptability across datasets. Visual comparisons are also given in Figure 8. Compared to other methods, our method authentically restores both accurate illumination and textual details.

Methods	PSNR ↑	SSIM ↑	LPIPS ↓
SCI (Ma et al. 2022)	17.30	0.540	0.308
KinD++ (Zhang et al. 2021)	17.66	0.761	0.217
Zero-DCE (Guo et al. 2020)	18.06	0.580	0.313
URetinex-Net (Wu et al. 2022)	21.22	0.859	0.099
IAT (Cui et al. 2022)	26.46	0.843	0.180
LLFormer (Wang et al. 2023b)	27.75	0.860	0.143
LLFlow (Wang et al. 2022c)	28.34	0.920	0.076
HWMNet (Fan, Liu, and Liu 2022)	30.30	0.910	0.080
MIRNetv2 (Zamir et al. 2022)	30.88	0.902	0.090
SNR (Xu et al. 2022)	30.92	0.894	0.139
Pydiff (Zhou, Yang, and Yang 2023)	<u>31.11</u>	<u>0.922</u>	<u>0.069</u>
MDMS (ours)	**33.30**	**0.933**	**0.043**

Table 2: Quantitative results on the LOLv2-Real dataset in terms of PSNR, SSIM and LPIPS. ↑ means higher is better, while ↓ means lower is better. The best performance is marked in bold with the second performance underlined.

Figure 8: Qualitative comparison with existing methods on LOLv2 dataset. The top row shows the results of synthesized images from LOLv2-Syn, while the bottom row shows real-world results from LOLv2-Real. Please zoom in to see the details.

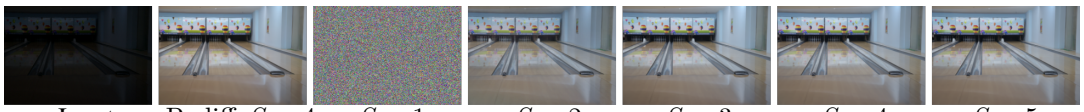

Figure 9: Results under different total sampling step S. Our method achieves promising results within three sampling steps.

Methods	PSNR ↑	SSIM ↑	LPIPS ↓
IAT (Cui et al. 2022)	15.37	0.710	0.279
SCI (Ma et al. 2022)	15.43	0.744	0.233
HWMNet (Fan, Liu, and Liu 2022)	15.76	0.743	0.252
SNR (Xu et al. 2022)	16.11	0.747	0.293
MIRNetv2 (Zamir et al. 2022)	16.38	0.786	0.245
LLFormer (Wang et al. 2023b)	17.16	0.784	0.244
Pydiff (Zhou, Yang, and Yang 2023)	17.33	**0.797**	0.255
MDMS (ours)	**17.40**	**0.797**	**0.227**

Table 3: Quantitative results on the LOLv2-Syn dataset.

Ablation	Variants	PSNR ↑	SSIM ↑	LPIPS ↓
MDL	w/o spatial	26.36	0.871	0.091
	w/o frequency	26.46	0.876	0.087
MSS	patch 64	26.06	0.876	0.085
	patch 96	26.40	0.850	0.105
	patch 128	25.99	0.815	0.148
	patch 64+96	26.78	0.880	0.081
BCP	w/o BCP	26.80	0.875	0.094
Full model	MDL + MSS + BCP	**27.12**	**0.882**	**0.078**

Table 4: Ablation studies on our MDL, MSS and BCP.

Ablation Study

We conduct several ablation studies to validate the effectiveness of our proposed Multi-Domain Learning (MDL), Multi-Scale Sampling (MSS), and Bright Channel Prior (BCP). All quantitative results as listed in Table 4. **(1):** For MDL, we individually remove the spatial branch and the frequency branch, then retrain the network to compare their performance. It is evident that the absence of any branch detrimentally impacts the final performance. **(2):** For MSS, we employ different sizes of patches during the sampling process. As shown in Figure 9 and Table 4, our multi-scale patch strategy effectively expands the sampling pool, which makes MDMS require fewer steps (total sampling step $S = 3$) to achieve higher performance than previous diffusion models. Visual comparisons in Figure 10 also show that MSS achieves the best trade-off between image smoothness and details. **(3):** For BCP, we replace the prior with Histogram Equalization for comparison. The performance in Table 4 proves the superiority of our bright channel prior.

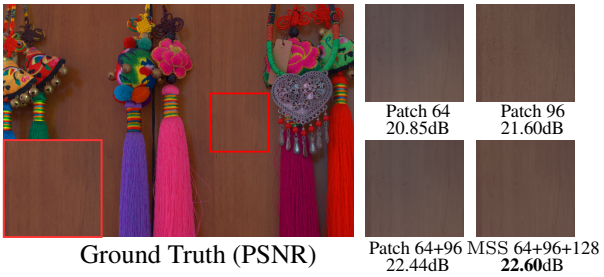

Figure 10: Ablation study on the proposed Multi-Scale Sampling (MSS) strategy. Please zoom in to see the details.

Conclusion

This paper proposes a Multi-Domain Multi-Scale (MDMS) diffusion-based method for low-light image enhancement. MDMS introduces a novel multi-domain learning paradigm, which explicitly captures more detailed features from an additional frequency domain. Furthermore, MDMS uses a multi-scale sampling strategy to alleviate checkerboard artifacts caused by uneven overlapping, and significantly improves performance with a more natural illumination guidance called bright channel prior. Extensive experiments on three benchmarks show that MDMS significantly outperforms other state-of-the-art low-light enhancement methods.

Acknowledgments

This work is supported by National Key Research and development Program of China (2021YFA1000102), National Natural Science Foundation of China (Nos. 62376285, 61673396) and Natural Science Foundation of Shandong Province, China (No. ZR2022MF260).

References

Cui, Z.; Li, K.; Gu, L.; Su, S.; Gao, P.; Jiang, Z.; Qiao, Y.; and Harada, T. 2022. Illumination adaptive transformer. *arXiv preprint arXiv:2205.14871*.

Dhananjaya, M. M.; Kumar, V. R.; and Yogamani, S. 2021. Weather and light level classification for autonomous driving: Dataset, baseline and active learning. In *2021 IEEE International Intelligent Transportation Systems Conference (ITSC)*, 2816–2821. IEEE.

Dhariwal, P.; and Nichol, A. 2021. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34: 8780–8794.

Eriksson, O. 2020. Real-world low-light image enhancement using Variational Autoencoders. *Master's Theses in Mathematical Sciences*.

Fan, C.-M.; Liu, T.-J.; and Liu, K.-H. 2022. Half wavelet attention on M-Net+ for low-light image enhancement. In *2022 IEEE International Conference on Image Processing (ICIP)*, 3878–3882. IEEE.

Graps, A. 1995. An introduction to wavelets. *IEEE computational science and engineering*, 2(2): 50–61.

Guo, C.; Li, C.; Guo, J.; Loy, C. C.; Hou, J.; Kwong, S.; and Cong, R. 2020. Zero-reference deep curve estimation for low-light image enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1780–1789.

Guo, X. 2016. LIME: A method for low-light image enhancement. In *Proceedings of the 24th ACM international conference on Multimedia*, 87–91.

He, K.; Sun, J.; and Tang, X. 2010. Single image haze removal using dark channel prior. *IEEE transactions on pattern analysis and machine intelligence*, 33(12): 2341–2353.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851.

Jiang, H.; Luo, A.; Han, S.; Fan, H.; and Liu, S. 2023. Low-Light Image Enhancement with Wavelet-based Diffusion Models. *arXiv preprint arXiv:2306.00306*.

Jiang, Y.; Gong, X.; Liu, D.; Cheng, Y.; Fang, C.; Shen, X.; Yang, J.; Zhou, P.; and Wang, Z. 2021. Enlightengan: Deep light enhancement without paired supervision. *IEEE transactions on image processing*, 30: 2340–2349.

Jobson, D. J.; Rahman, Z.-u.; and Woodell, G. A. 1997. Properties and performance of a center/surround retinex. *IEEE transactions on image processing*, 6(3): 451–462.

Kaur, M.; Kaur, J.; and Kaur, J. 2011. Survey of contrast enhancement techniques based on histogram equalization. *International Journal of Advanced Computer Science and Applications*, 2(7).

LeCun, Y.; Bengio, Y.; and Hinton, G. 2015. Deep learning. *nature*, 521(7553): 436–444.

Lin, M.; Chen, Q.; and Yan, S. 2013. Network in network. *arXiv preprint arXiv:1312.4400*.

Liu, L.; Liu, J.; Yuan, S.; Slabaugh, G.; Leonardis, A.; Zhou, W.; and Tian, Q. 2020. Wavelet-based dual-branch network for image demoiréing. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIII 16*, 86–102. Springer.

Liu, R.; Ma, L.; Zhang, J.; Fan, X.; and Luo, Z. 2021. Retinex-inspired unrolling with cooperative prior architecture search for low-light image enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10561–10570.

Lu, K.; and Zhang, L. 2020. TBEFN: A two-branch exposure-fusion network for low-light image enhancement. *IEEE Transactions on Multimedia*, 23: 4093–4105.

Luo, Z.; Gustafsson, F. K.; Zhao, Z.; Sjölund, J.; and Schön, T. B. 2023. Image restoration with mean-reverting stochastic differential equations. *arXiv preprint arXiv:2301.11699*.

Lv, F.; Lu, F.; Wu, J.; and Lim, C. 2018. MBLLEN: Low-Light Image/Video Enhancement Using CNNs. In *BMVC*, volume 220, 4.

Ma, L.; Ma, T.; Liu, R.; Fan, X.; and Luo, Z. 2022. Toward fast, flexible, and robust low-light image enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5637–5646.

Özdenizci, O.; and Legenstein, R. 2023. Restoring vision in adverse weather conditions with patch-based denoising diffusion models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Phung, H.; Dao, Q.; and Tran, A. 2023. Wavelet diffusion models are fast and scalable image generators. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10199–10208.

Rahman, Z.-u.; Jobson, D. J.; and Woodell, G. A. 2004. Retinex processing for automatic image enhancement. *Journal of Electronic imaging*, 13(1): 100–110.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2021. High-Resolution Image Synthesis with Latent Diffusion Models. arXiv:2112.10752.

Shao, M.; Qiao, Y.; Meng, D.; and Zuo, W. 2023. Uncertainty-guided hierarchical frequency domain Transformer for image restoration. *Knowledge-Based Systems*, 263: 110306.

Song, J.; Meng, C.; and Ermon, S. 2020. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*.

Wang, C.; Zheng, Z.; Quan, R.; Sun, Y.; and Yang, Y. 2023a. Context-Aware Pretraining for Efficient Blind Image Decomposition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18186–18195.

Wang, H.; Chen, Y.; Cai, Y.; Chen, L.; Li, Y.; Sotelo, M. A.; and Li, Z. 2022a. SFNet-N: An improved SFNet algorithm for semantic segmentation of low-light autonomous driving road scenes. *IEEE Transactions on Intelligent Transportation Systems*, 23(11): 21405–21417.

Wang, L.-W.; Liu, Z.-S.; Siu, W.-C.; and Lun, D. P. 2020. Lightening network for low-light image enhancement. *IEEE Transactions on Image Processing*, 29: 7984–7996.

Wang, T.; Zhang, K.; Shen, T.; Luo, W.; Stenger, B.; and Lu, T. 2023b. Ultra-high-definition low-light image enhancement: a benchmark and transformer-based method. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 2654–2662.

Wang, W.; Wang, X.; Yang, W.; and Liu, J. 2022b. Unsupervised face detection in the dark. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1): 1250–1266.

Wang, W.; Wei, C.; Yang, W.; and Liu, J. 2018. Gladnet: Low-light enhancement network with global awareness. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, 751–755. IEEE.

Wang, Y.; Wan, R.; Yang, W.; Li, H.; Chau, L.-P.; and Kot, A. 2022c. Low-light image enhancement with normalizing flow. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, 2604–2612.

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.

Wang, Z.; Jiang, Y.; Zheng, H.; Wang, P.; He, P.; Wang, Z.; Chen, W.; and Zhou, M. 2023c. Patch diffusion: Faster and more data-efficient training of diffusion models. *arXiv preprint arXiv:2304.12526*.

Wei, C.; Wang, W.; Yang, W.; and Liu, J. 2018. Deep retinex decomposition for low-light enhancement. *arXiv preprint arXiv:1808.04560*.

Wu, W.; Weng, J.; Zhang, P.; Wang, X.; Yang, W.; and Jiang, J. 2022. Uretinex-net: Retinex-based deep unfolding network for low-light image enhancement. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5901–5910.

Xu, X.; Wang, R.; Fu, C.-W.; and Jia, J. 2022. SNR-aware low-light image enhancement. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 17714–17724.

Yan, Y.; Ren, W.; Guo, Y.; Wang, R.; and Cao, X. 2017. Image deblurring via extreme channels prior. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4003–4011.

Yang, L.; Zhang, Z.; Song, Y.; Hong, S.; Xu, R.; Zhao, Y.; Shao, Y.; Zhang, W.; Cui, B.; and Yang, M.-H. 2022. Diffusion models: A comprehensive survey of methods and applications. *arXiv preprint arXiv:2209.00796*.

Yang, W.; Wang, S.; Fang, Y.; Wang, Y.; and Liu, J. 2020. From fidelity to perceptual quality: A semi-supervised approach for low-light image enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3063–3072.

Yang, W.; Wang, W.; Huang, H.; Wang, S.; and Liu, J. 2021. Sparse gradient regularized deep retinex network for robust low-light image enhancement. *IEEE Transactions on Image Processing*, 30: 2072–2086.

Zamir, S. W.; Arora, A.; Khan, S.; Hayat, M.; Khan, F. S.; Yang, M.-H.; and Shao, L. 2022. Learning enriched features for fast image restoration and enhancement. *IEEE transactions on pattern analysis and machine intelligence*, 45(2): 1934–1948.

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.

Zhang, Y.; Guo, X.; Ma, J.; Liu, W.; and Zhang, J. 2021. Beyond brightening low-light images. *International Journal of Computer Vision*, 129(4): 1013–1037.

Zhou, D.; Yang, Z.; and Yang, Y. 2023. Pyramid Diffusion Models For Low-light Image Enhancement. *arXiv preprint arXiv:2305.10028*.

Zhou, M.; Yu, H.; Huang, J.; Zhao, F.; Gu, J.; Loy, C. C.; Meng, D.; and Li, C. 2022. Deep fourier up-sampling. *arXiv preprint arXiv:2210.05171*.