Diff-Shadow: Global-guided Diffusion Model for Shadow Removal

Jinting Luo[1], Ru Li[2], Chengzhi Jiang[1], Xiaoming Zhang[3],
Mingyan Han[1], Ting Jiang[1], Haoqiang Fan[1], Shuaicheng Liu[4*]

[1]Megvii Technology Inc.
[2]Harbin Institute of Technology
[3]Southwest Jiaotong University
[4]University of Electronic Science and Technology of China
{k531756653, morven126, zxiaoming360, hanmy628, tjhedlen}@gmail.com,
liru@hit.edu.cn, fhq@megvii.com, liushuaicheng@uestc.edu.cn

Abstract

We propose Diff-Shadow, a global-guided diffusion model for shadow removal. Previous transformer-based approaches can utilize global information to relate shadow and non-shadow regions but are limited in their synthesis ability and recover images with obvious boundaries. In contrast, diffusion-based methods can generate better content but they are not exempt from issues related to inconsistent illumination. In this work, we combine the advantages of diffusion models and global guidance to achieve shadow-free restoration. Specifically, we propose a parallel UNets architecture: 1) the local branch performs the patch-based noise estimation in the diffusion process, and 2) the global branch recovers the low-resolution shadow-free images. A Reweight Cross Attention (RCA) module is designed to integrate global contextual information of non-shadow regions into the local branch. We further design a Global-guided Sampling Strategy (GSS) that mitigates patch boundary issues and ensures consistent illumination across shaded and unshaded regions in the recovered image. Comprehensive experiments on datasets ISTD, ISTD+, and SRD have demonstrated the effectiveness of Diff-Shadow. Compared to state-of-the-art methods, our method achieves a significant improvement in terms of PSNR, increasing from 32.33dB to 33.69dB on the ISTD dataset.

Introduction

Shadows are a phenomenon that exists when an optical image is captured with light blocked. The presence of shadow can complicate image processing as well as many subsequent vision tasks, e.g., object detection, tracking, and semantic segmentation (Guo et al. 2023a). The goal of shadow removal is to enhance the brightness of the image shadow regions and to achieve a consistent illumination distribution between shadow and non-shadow regions. The shadow removal task is essentially an image restoration task designed to utilize the information available in the shadow-affected image to recover the information lost due to light occlusion.

Shadow removal has received intensive and long-lasting attention. Traditional approaches use hand-crafted features and are based on physical lighting models (Gryka, Terry, and Brostow 2015; Zhang, Zhang, and Xiao 2015; Xiao et al.
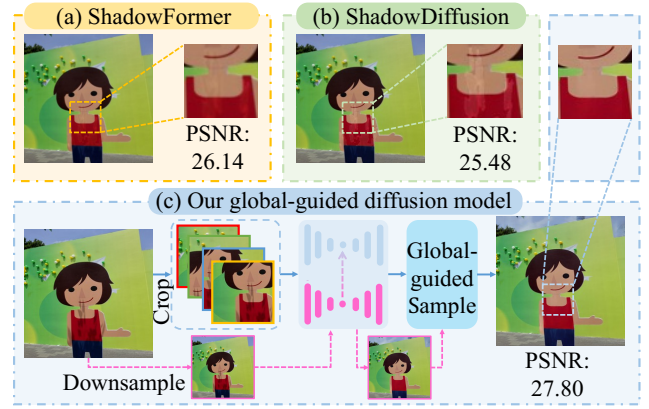
Figure 1: (a) shows the result of ShadowFormer (Guo et al. 2023a), which suffers from residual shadow artifacts due to the limited modeling ability, while the result of (b) ShadowDiffusion (Guo et al. 2023b) shows obvious illumination inconsistency across the images because it cannot exploit the global information. (c) exhibits the proposed Diff-Shadow, which generates high-quality shadow removal results that maintain illumination consistency and are free from block boundary artifacts through the design of a parallel network structure and a novel global-guided sampling strategy.

2013). Inspired by the successes of deep learning, many approaches that utilize large amounts of data (Guo et al. 2023a; Chen et al. 2021; Fu et al. 2021; Zhu et al. 2022a) have been introduced to improve shadow removal and reveal promising improvements, *e.g.*, Guo *et al.* proposed a transformer-based architecture, named ShadowFormer, to exploit non-shadow information to help shadow region restoration. Their shadow-interaction module plays an important role in exploring global information. However, this method often overlooks the modeling of the underlying distribution of natural images. Consequently, the generated results appear to have significant residual shadows, as shown in Fig. 1 (a).

Recently, diffusion models have gained wide interest in the field of image restoration (Saharia et al. 2022a,b; Jiang et al. 2023) due to its powerful modeling ability. Guo *et al.* proposed ShadowDiffusion to jointly pursue shadow-free images and refined shadow masks (Guo et al. 2023b).

Although it covers the whole input image, the kind of

conventional diffusion-based methods use a simple UNet and are challenging for shadow regions to capture effective information from matched non-shadow regions, resulting in illumination inconsistency in Fig. 1 (b). Furthermore, generating arbitrary resolution images directly with ShadowDiffusion proved challenging as diffusion models are more suitable for processing smaller patches (Ding et al. 2023). While Zheng *et al.* (Ding et al. 2023) and Ozan *et al.* (Özdenizci and Legenstein 2023) proposed several patch-based diffusion methods that demonstrated exceptional performance in image restoration tasks at arbitrary resolutions, they still faced limitations when handling shadow removal task. These methods lack global contextual information and ultimately lead to lighting inconsistencies or obvious boundaries in the restored images.

Considering the powerful modeling capabilities of diffusion models and the necessity of global information for shadow removal tasks, we propose Diff-Shadow, a global-guided diffusion model for generating high-quality shadow-free results with no obvious boundaries and consistent illumination in arbitrary image resolution. Our Diff-Shadow adopts the parallel UNets network architecture that receives both patch and global inputs, where the local branch focuses on estimating patch noise, while the global branch is responsible for recovering low-resolution shadow-free images. Additionally, our Diff-Shadow exploits the global contextual information of the non-shadow region from the global branch through the Reweight Cross Attention (RCA) module. Subsequently, we present a Global-guided Sampling Strategy (GSS) within the denoising process. In this strategy, the fusion weights for patch noise are determined by considering both the brightness disparity between the recovered patch and the global image, as well as the extent of shadow regions within the patch. This approach addresses potential merging artifacts, preserves superb illumination coherence across shaded and non-shadow regions, and ultimately enhances the overall quality of the results, as shown in Fig. 1 (c). Moreover, benefiting from the patch-based methods, the Diff-Shadow can be expanded to handle images with arbitrary resolution.

To summarize, our main contributions are as follows:

(1) We propose Diff-Shadow, a global-guided diffusion architecture that incorporates parallel UNets and exploits the correlation between the local branch and the non-shadow regions of the global through a reweighted cross-attention (RCA) module, which removes shadows efficiently.

(2) We propose a Global-guided Sampling Strategy (GSS) to further improve the quality of shadow removal during the diffusion denoising process.

(3) Extensive experimental results on public ISTD (Wang, Li, and Yang 2018), ISTD+ (Le and Samaras 2019), and SRD (Qu et al. 2017) datasets demonstrate that the proposed Diff-Shadow outperforms the state of the art.

Code — https://github.com/MonteCarluo/Diff-Shadow

Related Works

Shadow Removal. Recent learning-based methods for shadow removal tasks have achieved superior performance by leveraging large-scale high-quality training data to learn high-level image texture and semantic information (Chen et al. 2021; Fu et al. 2021; Guo et al. 2023a,b). Unlike traditional physical-based methods that rely on prior assumptions like illumination (Arbel and Hel-Or 2010; Zhang, Zhang, and Xiao 2015), gradient (Finlayson, Hordley, and Drew 2002; Finlayson et al. 2005; Mohan, Tumblin, and Choudhury 2007; Liu and Gleicher 2008; Gryka, Terry, and Brostow 2015), and region (Guo, Dai, and Hoiem 2012; Guo et al. 2021, 2022), the learning-based approaches tend to learn contextual features for shadow removal. For instance, Qu *et al.* proposed a multi-context embedding network to integrate multi-level information (Qu et al. 2017), while Wang *et al.* explored the relationship between shadow detection and removal, and proposed a stacked conditional generative adversarial network (ST-CGAN) (Wang, Li, and Yang 2018). Zhu *et al.* proposed a model-driven network (Zhu et al. 2022b) combining efficiency and interpretability by formulating a new shadow illumination model and unfolding it into a deep network. Other works like Hu *et al.* (Hu et al. 2019) and Zhang *et al.* (Zhang et al. 2020) focused on applying spatial context attention features and exploring residual-illumination relationships, respectively. These learning-based methods have produced impressive results, but there remain opportunities to further capture the correspondence between shadow and non-shadow areas.

Shadow and Non-Shadow Consistency. Shadow removal presents a unique challenge in that shadows corrupt only parts of an image, while abundant uncorrupted information remains in non-shadowed regions. However, previous methods (Le and Samaras 2019, 2020) focused only on local shadow or non-shadow regions, leading to severe artifacts around shadow boundaries and inconsistent brightness between shadow and non-shadow areas. More recent approaches leverage contextual information from both shadow and non-shadow areas of the image to inform the shadow removal process. For instance, DeshadowNet (Qu et al. 2017) incorporates multi-scale information to expand the receptive field, leveraging texture and appearance information across the image. Chen *et al.* proposed CANet (Chen et al. 2021), using an external patch-matching module to discover texture relationships. Jin *et al.* proposed SG-ShadowNet (Wan et al. 2022), reformulating shadow removal as an intra-image style transfer problem to harmonize de-shadowed regions with non-shadow areas. Guo *et al.* introduced ShadowFormer (Guo et al. 2023a), incorporating a Shadow-Interaction Module in the transformer to model the global texture. Related but divergent, we propose the Reweight Cross Attention module injected into the diffusion model, empowering it with global textural relationship modeling.

Method

Given the intermediate variable $\mathbf{x}_t \in \mathbb{R}^{3 \times H \times W}$, the shadow image $\tilde{\mathbf{x}} \in \mathbb{R}^{3 \times H \times W}$ with corresponding shadow mask $\mathbf{x}_m \in \mathbb{R}^{1 \times H \times W}$ in Fig. 2 (a), the proposed architecture is trained and inference using both high-resolution image patches as well as low-resolution down-sampled full images, which are optimized by the parallel UNets (Fig. 2 (d)) whose local branch performs the diffusion noise estimation of local
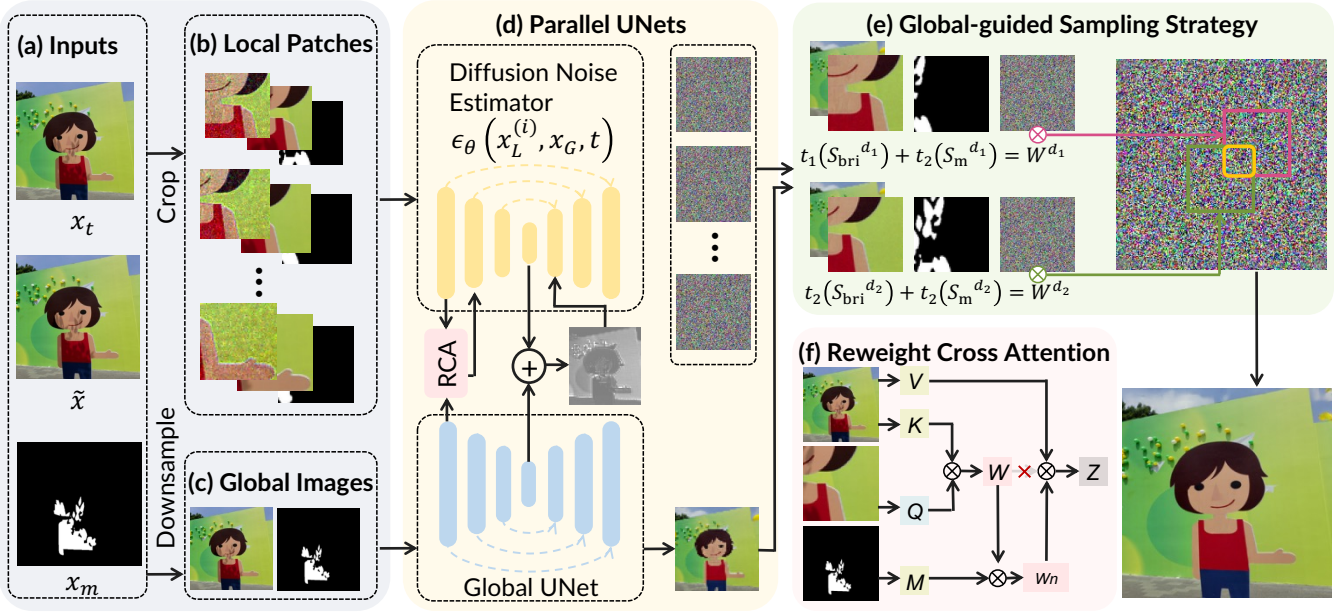
Figure 2: An overview of the Diff-Shadow. The local branch of Parallel UNets performs patch diffusion noise estimation using patches of the intermediate variable $\mathbf{x}_t$, the shadow image $\tilde{\mathbf{x}}$, and the shadow mask $\mathbf{x}_m$. The latter two images are also down-sampled as inputs of the global UNet, which constructs the low-resolution shadow-free image on the one hand and provides the global contextual information of non-shaded regions $\mathbf{x}_G$ into the local branch $\mathbf{x}_L^{(i)}$ using the proposed Reweight Cross Attention (RCA) module on the other hand. After estimating the noises of local patches, the Global-guided Sampling Strategy (GSS) works to construct the noise distribution for the whole image. t_1 and t_2 represent functions corresponding to the step t.

patches and the global UNet recovers the down-sampled full shadow-free image. The global information of non-shadow regions is integrated into the local branch through the RCA module (Fig. 2 (f)) and the Convolutional Aggregation (CA) during training. The GSS (Fig. 2 (e)) is utilized to combine the features of neighboring patches during inference.

Overall Architecture

Condition diffusion model (Saharia et al. 2022b) have achieved impressive performance for image-conditional data synthesis and editing tasks, which learns a conditional reverse process $p_\theta (\mathbf{x}_{0:T} \mid \tilde{\mathbf{x}})$ without modifying the diffusion process $q (\mathbf{x}_{1:T} \mid \mathbf{x}_0)$ for $\mathbf{x}$, such that the sampled image has high fidelity to the data distribution conditioned on $\tilde{\mathbf{x}}$. Common diffusive models train samples $(\mathbf{x}_0, \tilde{\mathbf{x}}) \sim q (\mathbf{x}_0, \tilde{\mathbf{x}})$ from a paired data distributions (e.g., a clean image $\mathbf{x}_0$ and a degraded image $\tilde{\mathbf{x}}$), and learn a conditional diffusion model that provides $\tilde{\mathbf{x}}$ to the reverse process:

$$p_\theta (\mathbf{x}_{0:T} \mid \tilde{\mathbf{x}}) = p (\mathbf{x}_T) \prod_{t=1}^{T} p_\theta (\mathbf{x}_{t-1} \mid \mathbf{x}_t, \tilde{\mathbf{x}}). \quad (1)$$

As for shadow removal tasks with shadow masks $\mathbf{x}_m$, we design the parallel UNets architecture. The local branch performs the diffusion noise estimation and is conditioned by image patches: $\mathbf{x}_L^{(i)} = \{\tilde{\mathbf{x}}^{(i)}, \mathbf{x}_m^{(i)}\}$, where $\tilde{\mathbf{x}}^{(i)} = \mathrm{Crop} (\boldsymbol{L}_i \circ \tilde{\mathbf{x}})$ and $\mathbf{x}_m^{(i)} = \mathrm{Crop} (\boldsymbol{L}_i \circ \mathbf{x}_m)$ represent $R \times R$ patches, and $\mathrm{Crop}(\cdot)$ extracts the patches from the location $\boldsymbol{L}_i$. The estimated noise of local patches will be com-

bined during the inference through the proposed GSS module (Fig. 2 (e)). The global branch is optimized by the low-resolution down-sampled full images: $\mathbf{x}_G = \{\tilde{\mathbf{x}}', \mathbf{x}_m'\}$. The contextual information of non-shadow regions plays an auxiliary role and is integrated into the local branch through the RCA module (Fig. 2 (f)). Therefore, the reverse process in Eq. 1 can be expanded as:

$$p_\theta \left(\mathbf{x}_{0:T}^{(i)} \mid \mathbf{x}_L^{(i)}, \mathbf{x}_G\right) = p \left(\mathbf{x}_T^{(i)}\right) \prod_{t=1}^{T} p_\theta \left(\mathbf{x}_{t-1}^{(i)} \mid \mathbf{x}_t^{(i)}, \mathbf{x}_L^{(i)}, \mathbf{x}_G\right),$$

$$(2)$$

where $\mathbf{x}_0^{(i)} = \mathrm{Crop} (\boldsymbol{L}_i \circ \mathbf{x}_0)$. The training approach is outlined in Algorithm 1.

We can sample the intermediate $\mathbf{x}_t$ from the clean image $\mathbf{x}_0$ through $\mathbf{x}_t = \sqrt{\bar{\alpha}_t} \mathbf{x}_0 + \sqrt{1 - \bar{\alpha}_t} \boldsymbol{\epsilon}$, where $\alpha_t = 1 - \beta_t$, β_t is a predefined variance schedule, $\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i$, and $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. The model is trained to predict the noise by:

$$\boldsymbol{\epsilon}_t = \boldsymbol{\epsilon}_\theta \left(\sqrt{\bar{\alpha}_t} \mathbf{x}_0^{(i)} + \sqrt{1 - \bar{\alpha}_t} \boldsymbol{\epsilon}, \mathbf{x}_L^{(i)}, \mathbf{x}_G, t\right). \quad (3)$$

Reweight Cross Attention

Existing patch-based methods lack generality since the regions inside a patch may all be corrupted. Apart from the local patch-based diffusion branch, we introduce a parallel global UNet to recover low-resolution shadow-free images. The Reweight Cross Attention (RCA) module is introduced to integrate the global information into the local patch by combining the low-level features. As shown in Fig. 2 (f), the

Input: shadow image $\tilde{\mathbf{x}}$, clean image $\mathbf{x}_0$ and mask $\mathbf{x}_m$
1: **repeat**
2:     Randomly sample a patch $\boldsymbol{L}_i$
3:     $\mathbf{x}_0^{(i)} = \text{Crop}\left(\boldsymbol{L}_i \circ \mathbf{x}_0\right)$
4:     $\mathbf{x}_L^{(i)} = \left\{\tilde{\mathbf{x}}^{(i)}, \mathbf{x}_m^{(i)}\right\}$ where $\tilde{\mathbf{x}}^{(i)} = \text{Crop}\left(\boldsymbol{L}_i \circ \tilde{\mathbf{x}}\right)$ and
       $\mathbf{x}_m^{(i)} = \text{Crop}\left(\boldsymbol{L}_i \circ \mathbf{x}_m\right)$
5:     $\mathbf{x}_G = \{\tilde{\mathbf{x}}', \mathbf{x}_m'\}$ where $\tilde{\mathbf{x}}' = \text{Dowmsample}(\tilde{\mathbf{x}})$ and
       $\mathbf{x}_m' = \text{Dowmsample}(\mathbf{x}_m)$
6:     $t \sim \text{Uniform}\{1, \ldots, T\}$
7:     $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
8:     Perform a gradient descent step for
$$\nabla_\theta \left\| \boldsymbol{\epsilon}_t - \boldsymbol{\epsilon}_\theta\left(\sqrt{\bar{\alpha}_t}\mathbf{x}_0^{(i)} + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, \mathbf{x}_L^{(i)}, \mathbf{x}_G, t\right) \right\|^2$$
9: **until** converged
10: **return** θ

Figure 3: The illustration of noise merge for overlapping patches. N_p represents the number of overlapping patches for each pixel. r and R are the step size and the patch size.

output of RCA is:

$$\mathbf{Z} = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{d}\right)\mathbf{M}\mathbf{V}, \tag{4}$$

$\mathbf{Q}$ represents the projected queries of patch features, $\mathbf{K}$ and $\mathbf{V}$ are the projected keys and values of global image features, d is the scaling parameter, $\mathbf{M}$ is the binary mask to multiply with the features, and $\mathbf{Z}$ represents the output interacted features. Attention maps between $\mathbf{Q}$ and $\mathbf{K}$ are first calculated to capture the mutual information between pixels in the local branch and regions in the global branch. Subsequently, the mask is introduced to assist in suppressing mutual information values in the global shadow regions. The adjusted attention map is then multiplied by the global $\mathbf{V}$ vector to obtain the expected global information for the local branch.

We interact the low-level features of the first convolutional block in the global UNet with the corresponding layer in the local UNet through the RCA because low-level features concentrate more on image edges, colors, and corners, which are important for shadow removal tasks. As for high-level abstract features, we apply a simple concatenation to achieve interaction. In contrast to recent cross-attention modules in ShadowFormer (Guo et al. 2023a) whose $\mathbf{Q}$, $\mathbf{K}$ and $\mathbf{V}$ originate from the same global feature, the RCA integrates both local and global information. Furthermore, other methods boost the mutual information between shadow and non-shadow pairs through masks. However, the shadow information is dispensable for recovering non-shadow regions. Contrarily, our attention mechanism utilizes masks to suppress the global shadow values through the attention map, integrating the more important non-shadow information from the global branch into the local branch.

Global-guided Samping Strategy

To overcome the shortcomings of merging artifacts from independently intermediate patches, Ozdenizci *et al.* proposed to perform reverse sampling based on the mean estimated noise for each pixel in overlapping patch regions, which works better for weather image with relatively unified degradation (Özdenizci and Legenstein 2023). However, as for shadow removal tasks with extreme inconsistencies between

shadow and non-shadow regions, predicting mean estimated noise lacks ability to achieve a consistent illumination distribution. Therefore, we propose the GSS to maintain illumination consistency.

The general idea of patch-based restoration is to operate locally on patches extracted from the image and optimally merge the results. We consider a grid-like arranged parsing scheme over the intermediate variable $\mathbf{x}_t$, the shadow image $\tilde{\mathbf{x}}$ and the shadow mask $\mathbf{x}_m$, and extract all patches by moving over the $R \times R$ grid with a step size of r in both horizontal and vertical dimensions, as shown in Fig. 3. During patch-based diffusive image restoration, we perform noise estimation based on adaptive estimated noise for each pixel in overlapping patch regions, at any given denoising time step t. As shown in Fig. 2 (e), the weight of each patch is determined by two parts: the brightness-based score S_{bri} and the mask-based score S_{m}. The former one considers the brightness consistency of the local patches and the global images, while the latter one corresponds to the proportion of shadow and non-shadow regions.

The brightness-based score S_{bri} constrains that the patch recovered by the local branch should show similar brightness with the full image generated by the global branch. The S_{bri} is defined as:

$$S_{\text{bri}} = 1 - \frac{\text{avg}(|L_{\text{gray}} - G_{\text{gray}}|)}{\max((|L_{\text{gray}} - G_{\text{gray}}|))}, \tag{5}$$

where L_{gray} is the recovered gray image of the local patch, G_{gray} is the recovered gray image of global image.

In addition, the masked-based score S_{m} pays more attention to non-shadow regions. When the non-shadow region dominates the image patch, the receptive field of the local branch encompasses both shadow and non-shadow regions, providing a valuable way to eliminate shadows through non-shadow information. Conversely, when the non-shadow region dominates the image patch, the predicted illumination may mainly originate from the global branch, whose performance tends to be unsatisfactory. Based on the binary mask, the ratio of non-shadow region among the patch can be calculated through: $R_{\text{sf}} = P_{\text{sf}}/P_{\text{total}}$, where P_{sf} is the number of non-shadow pixels and P_{total} is the number of total patch pixels. The mask-based score S_{m} is defined as:

$$S_{\text{m}} = \begin{cases} 1 & R_{\text{sf}} \geq T_{\text{sf}} \\ R_{\text{sf}}/T_{\text{sf}} & R_{\text{sf}} < T_{\text{sf}} \end{cases}, \tag{6}$$

Algorithm 2: Global-guided diffusive image restoration

Input: shadow image $\tilde{\mathbf{x}}$, shadow mask $\mathbf{x}_m$, conditional diffusion model $\boldsymbol{\epsilon}_\theta$, number of implicit sampling steps S, dictionary of D overlapping patch locations.

1: $\mathbf{x}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
2: **for** $i = S, \ldots, 1$ **do**
3: $\quad t = (i-1) \cdot T/S + 1$
4: $\quad t_{\text{next}} = (i-2) \cdot T/S + 1$ if $i > 1$ else 0
5: $\quad \hat{\boldsymbol{\Omega}}_t = 0$ and $M = 0$
6: $\quad$ **for** $d = 1, \ldots, D$ **do**
7: $\qquad \mathbf{x}_t^{(d)} = \text{Crop}\left(\boldsymbol{L}_d \circ \mathbf{x}_t\right)$
8: $\qquad \mathbf{x}_L^{(d)} = \left\{\tilde{\mathbf{x}}^{(d)}, \mathbf{x}_m^{(d)}\right\}$ where $\tilde{\mathbf{x}}^{(d)} = \text{Crop}\left(\boldsymbol{L}_d \circ \tilde{\mathbf{x}}\right)$ and $\mathbf{x}_m^{(d)} = \text{Crop}\left(\boldsymbol{L}_d \circ \mathbf{x}_m\right)$
9: $\qquad \mathbf{x}_G = \left\{\tilde{\mathbf{x}}', \mathbf{x}_m'\right\}$ where $\mathbf{x}' = \text{Downsample}(\mathbf{x})$ and $\mathbf{x}_m' = \text{Downsample}(\mathbf{x}_m)$
10: $\qquad \hat{\boldsymbol{\Omega}}_t = \hat{\boldsymbol{\Omega}}_t + W^d \cdot \boldsymbol{\epsilon}_\theta\left(\mathbf{x}_t^{(d)}, \tilde{\mathbf{x}}_L^{(d)}, \mathbf{x}_G, t\right)$
11: $\qquad M = M + W^d$
12: $\quad$ **end for**
13: $\quad \hat{\boldsymbol{\Omega}}_t = \hat{\boldsymbol{\Omega}}_t / M$
14: $\quad \mathbf{x}_t \leftarrow \sqrt{\bar{\alpha}_{t_{\text{next}}}}\left(\frac{\mathbf{x}_t - \sqrt{1-\bar{\alpha}_t} \cdot \hat{\boldsymbol{\Omega}}_t}{\sqrt{\bar{\alpha}_t}}\right) + \sqrt{1 - \bar{\alpha}_{t_{\text{next}}}} \cdot \hat{\boldsymbol{\Omega}}_t$
15: **end for**
16: **return** $\mathbf{x}_t$

where T_{sf} is the threshold of the ratio. In our implementation, T_{sf} is set to 0.5 and eps is set to 0.01.

The weights of each patch contributed to constructing the full noise map are calculated as:

$$W^d = \frac{t \cdot S_{\text{bri}}}{T} + \frac{(T-t) \cdot S_{\text{m}}}{T}, \tag{7}$$

where T is the total steps. The final adaptive estimated noise is computed as:

$$\frac{1}{\sum_i W^{d_i}} \sum_{i=1}^{N_p} \left(W^{d_i} \cdot \boldsymbol{\epsilon}_\theta\left(\mathbf{x}_t^{(i)}, \mathbf{x}_L^{(i)}, \mathbf{x}_G, t\right)\right), \tag{8}$$

where $i \in \{1, \ldots, N_p\}$ and N_p is the number of overlapping patches of each pixel. The test time diffusive image restoration method is outlined in Algorithm 2.

Optimization

The Diff-Shadow involves two parallel UNet networks operating the local patch optimization and the global shadow-free image reconstruction. The objective function includes the following two items: the diffusive objective function L_{diff} and the global loss function L_{global}. Total loss $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{diff}} + \lambda \mathcal{L}_{\text{global}}$, where λ is a hyper-parameter to balance the contributions of the two losses. We empirically set it to 1.

The Diffusive Objective Function. Refer to (Ho, Jain, and Abbeel 2020), the $\mathcal{L}_{\text{diff}}$ is defined as:

$$\mathcal{L}_{\text{diff}} = \mathbb{E}_{\mathbf{x}_0, t, \boldsymbol{\epsilon}} \|\boldsymbol{\epsilon}_t - \boldsymbol{\epsilon}\|_F^2. \tag{9}$$

.

The Global Loss Function. The global UNet recovers shadow-free images from down-sampled shadow images and mask images:

$$\mathcal{L}_{\text{global}} = \mathbb{E}_{\mathbf{x_0}', \mathbf{x_m}'} \|\mathbf{x}_r - \mathbf{x_0}'\|_F^2, \tag{10}$$

where $\mathbf{x}_r$ is the global output, $\mathbf{x_0}'$ and $\mathbf{x_m}'$ are the down-sampled clean images and shadow mask, respectively.

Experiments

Experimental Setups

Datasets. We used three standard datasets: 1) ISTD dataset (Wang, Li, and Yang 2018) includes 1,330 training and 540 testing triplets (shadow images, masks and shadow-free images); 2) Adjusted ISTD (ISTD+) dataset (Le and Samaras 2019) reduces the illumination inconsistency between the shadow and shadow-free image of ISTD; 3) SRD dataset (Qu et al. 2017) consists of 2,680 training and 408 testing pairs of shadow and shadow-free images without the shadow masks. Referring to ShadowDiffusion, we applied the prediction mask provided by DHAN (Cun, Pun, and Shi 2020) for the training and testing.

Implementation Details. Our Diff-Shadow is trained using eight NVIDIA GTX 2080Ti GPUs. The Adam optimizer (Kingma and Ba 2015) is applied to optimize the parallel UNets with a fixed learning rate of $lr = 2e^{-4}$ without weight decay and the training epoch is set as 2,000. The exponential moving average (Nichol and Dhariwal 2021) with a weight of 0.999 is applied for parameter updating. Further details regarding the inference time and network complexity can be found in the supplementary.

Training Specifications. For each dataset, the images are resized into 256×256 for training. For each iteration, we initially sampled 8 images from the training set and randomly cropped 16 patches with size 64×64 from each image plus a down-sampled global image with the same size corresponding to each patch. This process resulted in mini-batches consisting of 96 samples. We used input time step embeddings for t through sinusoidal positional encoding (Vaswani et al. 2017) and provided these embeddings as inputs to each residual block in the local and global branches, allowing the model to share parameters across time. During sampling, the patch size R is set to 64 and the step size r is set to 8 for covering the whole image.

Evaluation Measures. Following previous works (Fu et al. 2021; Jin, Sharma, and Tan 2021; Zhu et al. 2022a; Guo et al. 2023a,b), we utilized the Root Mean Square Error (RMSE) in the LAB color space as the quantitative evaluation metric of the shadow removal results, comparing to the ground truth shadow-free images. Lower RMSE scores correspond to better performance. In addition, we also adopt the common Peak Signal-to-Noise Ratio (PSNR) and the Structural Similarity Index Measure (SSIM) (Wang et al. 2004) to measure the performance in the RGB color space.

Comparison with the State-of-the-art

We compare the proposed Diff-Shadow with the popular or state-of-the-art shadow removal algorithms, including one traditional method: Guo *et al.* (Guo, Dai, and Hoiem 2012), several deep learning-based methods: DeshadowNet (Qu et al. 2017), ST-CGAN (Wang, Li, and Yang 2018), DHAN (Cun, Pun, and Shi 2020), Fu *et al.* (Fu et al. 2021), DC-ShadowNet (Jin, Sharma, and Tan 2021), BM-Net (Zhu et al. 2022a), ShadowFormer (Guo et al. 2023a), and one diffusion-based method: ShadowDiffusion (Guo et al. 2023b). The shadow removal results of the competing methods are quoted from the original papers. Following

Dataset	Method	Shadow Regions (S)			Non-Shadow Regions (NS)			All Image (All)		
		PSNR↑	SSIM↑	RMSE↓	PSNR↑	SSIM↑	RMSE↓	PSNR↑	SSIM↑	RMSE↓
ISTD	Input Image	22.40	0.936	32.10	27.32	0.976	7.09	20.56	0.893	10.88
	Guo *et al.*	27.76	0.964	18.65	26.44	0.975	7.76	23.08	0.919	9.26
	ST-CGAN	33.74	0.981	9.99	29.51	0.958	6.05	27.44	0.929	6.65
	DHAN	35.53	0.988	7.49	31.05	0.971	5.30	29.11	0.954	5.66
	Fu *et al.*	34.71	0.975	7.91	28.61	0.880	5.51	27.19	0.945	5.88
	DC-ShadowNet	31.69	0.976	11.43	28.99	0.958	5.81	26.38	0.922	6.57
	Zhu et al.	36.95	0.987	8.29	31.54	0.978	4.55	29.85	0.960	5.09
	ShadowFormer	38.19	0.991	5.96	34.32	0.981	3.72	32.21	0.968	4.09
	ShadowDiffusion	**40.15**	**0.994**	4.13	33.70	0.977	4.14	32.33	0.969	4.12
	Ours	39.71	**0.994**	**1.41**	**35.58**	**0.987**	**2.48**	**33.69**	**0.976**	**2.89**
SRD	Input Image	18.96	0.871	36.69	31.47	0.975	4.83	18.19	0.830	14.05
	Guo *et al.*	-	-	29.89	-	-	6.47	-	-	12.60
	DeshadowNet	-	-	11.78	-	-	4.84	-	-	6.64
	DHAN	33.67	0.978	8.94	34.79	0.979	4.80	30.51	0.949	5.67
	Fu *et al.*	32.26	0.966	9.55	31.87	0.945	5.74	28.40	0.893	6.50
	DC-ShadowNet	34.00	0.975	7.70	35.53	0.981	3.65	31.53	0.955	4.65
	BMNet	35.05	0.981	6.61	36.02	0.982	3.61	31.69	0.956	4.46
	SG-ShadowNet	-	-	7.53	-	-	2.97	-	-	4.23
	Zhu et al.	34.94	0.980	7.44	35.85	0.982	3.74	31.72	0.952	4.79
	ShadowFormer	36.91	**0.989**	5.90	36.22	0.989	3.44	32.90	0.958	4.04
	ShadowDiffusion	**38.47**	0.987	4.98	37.78	0.985	3.44	34.73	0.957	3.63
	Ours	37.91	0.988	**1.81**	**39.49**	**0.990**	**1.65**	**34.93**	**0.977**	**2.52**
ISTD+	Input Image	18.96	0.871	36.69	31.47	0.975	4.83	18.19	0.830	14.05
	ST-CGAN	-	-	13.40	-	-	7.70	-	-	8.70
	DeshadowNet	-	-	15.90	-	-	6.00	-	-	7.60
	Fu *et al.*	36.04	0.976	6.60	31.16	0.876	3.83	29.45	0.840	4.20
	DC-ShadowNet	32.20	0.976	10.43	33.21	0.963	3.67	28.76	0.922	4.78
	BMNet	37.30	0.990	6.19	35.06	0.974	3.09	32.30	0.955	3.60
	SG-ShadowNet	-	-	5.9	-	-	2.90	-	-	3.4
	ShadowFormer	39.67	-	5.20	38.82	-	2.30	35.46	-	2.80
	ShadowDiffusion	39.82	-	4.90	38.90	-	2.30	35.68	0.970	2.70
	Ours	**40.63**	**0.993**	**1.30**	**39.69**	**0.989**	**1.57**	**36.47**	**0.979**	**2.09**

Table 1: The quantitative comparisons on ISTD, SRD, and ISTD+ datasets. '-' represents the values that were not reported.

the previous methods (Guo et al. 2023a,b), we evaluate the removal results with 256×256 resolution, while the results for other resolutions are detailed in the supplementary.

Quantitative Evaluation. Table 1 shows the quantitative results on the test sets over ISTD, SRD, and ISTD+, respectively. The superiority of the proposed method over competing approaches is evident across all three datasets, as it consistently achieves superior performance. Thanks to the powerful modeling capabilities of the diffusion model for images, the proposed method improves the PSNR of the SRD from 32.90 dB to 34.93 dB when compared with the transformer-based method ShadowFormer. Furthermore, compared to ShadowDiffusion which is also a diffusion-based method, our parallel network architecture interacts the local and global information more effectively. The novel GSS further enhances the accuracy of shadow-free image restoration, which results in an improvement in PSNR from 32.33 dB to 33.69 dB on the ISTD. In summary, our Diff-Shadow achieves new state-of-the-art results on three

widely-used benchmarks.

Qualitative Evaluation. We present the qualitative comparison in Fig. 4 and the supplementary. Diff-Shadow stands out among previous methods for its ability to remove shadows, while preserving the surrounding image textures and luminance, even in areas of high contrast that pose challenges for shadow removal, *e.g.*, in the third example of Fig. 4, there is a distinct color disparity between the ground and the shadow. Previous methods struggled to remove the fragmented shadows of branches at this location, resulting in uneven color distribution or incomplete removal results. The result of our method is the most similar to the ground truth.

Ablation Studies

Effectiveness of Parallel UNets. The global branch is removed to only accommodate the local inputs. Consequently, the network degrades to a single-head UNet structure. As in Table 2 (a), the single-head UNet shows the worst performance due to missing global context information. We tried
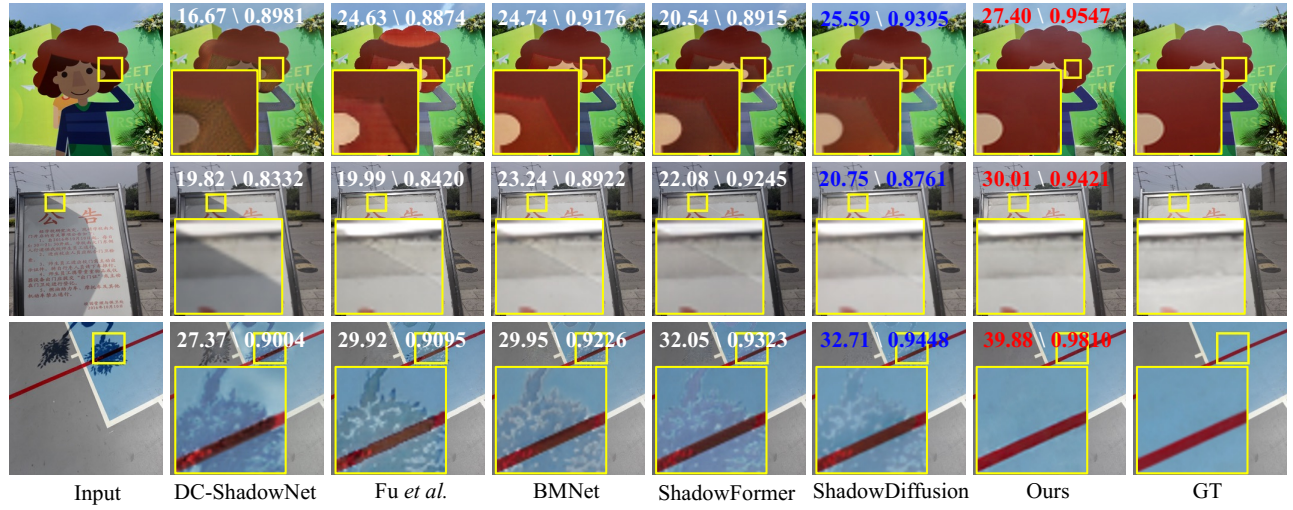
Figure 4: Qualitative comparisons on ISTD+. The values represent the 'PSNR/SSIM'.

Methods	S	NS	ALL
(a) Single-head(w/o global)	2.64	3.41	3.36
(b) Single-head(w global)	2.55	3.19	3.04
(c) Ours (w/o RCA)	2.31	2.14	2.73
(d) Ours (complete model)	**1.81**	**1.65**	**2.52**

Table 2: Ablation studies of different network structures on SRD dataset. We report RMSE values.

Sampling	S	NS	ALL
(a) w/o GSS	2.97	3.63	3.81
(b) $W^d = 1$	2.64	3.42	3.27
(c) S_{bri}	2.16	2.99	2.96
(d) $S_{\text{bri}} + S_{\text{m}}$	**1.81**	**1.65**	**2.52**

Table 3: Ablation studies of different sampling strategies on SRD dataset. We report RMSE values.
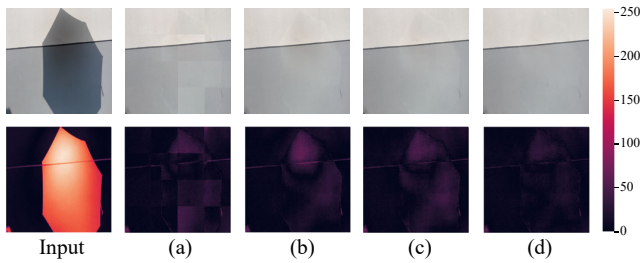

Figure 5: Visual examples of results and error maps for ablation study with different sampling strategies, along with corresponding results from Table 3.

to concatenate the patch and global images together at the network input level, and the results in Table 2 (b) show that the introduction of global context information works bet-

ter than patch input alone. Then, to verify the effectiveness of the RCA module, we retained the two-branch network mechanism but eliminated the RCA module from the first layer and kept only the concatenate connection in the middle layer. As shown in Table 2 (c), the RCA module is effective in interacting two branches and enhancing the final result.

Effectiveness of Global-guided Sampling Strategy. Table 3 (a) presents the experiments without overlapping between patches. It turns out significant boundaries between patches in Fig. 5 (a). Second, we established the adaptive noise weights as fixed $W^d = 1$ in Table 3 (b), which leads to a degradation of the sampling procedure. Reflected in Fig. 5 (b), we can see that there is a large discrepancy between the results and the ground truth. Third, we controlled the noise merge weights solely by the difference between the recovered patch brightness and the recovered global image brightness, *i.e.*, only brightness-based score S_{bri} exists. The qualitative patch image restored at the beginning of the sampling phase is heavily noisy in Fig. 5 (c). Therefore, it is not accurate to control the weights using brightness alone. The results in Table 3 (c) are inferior to those of our full strategy in Table 3 (d), which considers both the proportion of shaded areas in the patch and the brightness difference between the restored global and local image.

Conclusion

In this work, we propose a global-guided diffusion model named Diff-Shadow, for high-quality shadow removal with no obvious boundaries and consistent illumination. We design a parallel network structure in the framework of the diffusion model to assist in the recovery of local patches using the shadow-free regions from the global image. In addition, we propose a sampling strategy named GSS to further improve the performance of shadow removal during diffusion denoising. Comprehensive experiments demonstrate that Diff-shadow outperforms the state-of-the-art.

Acknowledgments

References

Arbel, E.; and Hel-Or, H. 2010. Shadow removal using intensity surfaces and texture anchor points. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(6): 1202–1216.

Chen, Z.; Long, C.; Zhang, L.; and Xiao, C. 2021. Canet: A context-aware network for shadow removal. In *International Conference on Computer Vision*, 4743–4752.

Cun, X.; Pun, C.-M.; and Shi, C. 2020. Towards ghost-free shadow removal via dual hierarchical aggregation network and shadow matting gan. In *AAAI Conference on Artificial Intelligence*, 10680–10687.

Ding, Z.; Zhang, M.; Wu, J.; and Tu, Z. 2023. Patched denoising diffusion models for high-resolution image synthesis. In *The Twelfth International Conference on Learning Representations*.

Finlayson, G. D.; Hordley, S. D.; and Drew, M. S. 2002. Removing shadows from images. In *European Conference on Computer Vision*, 823–836.

Finlayson, G. D.; Hordley, S. D.; Lu, C.; and Drew, M. S. 2005. On the removal of shadows from images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(1): 59–68.

Fu, L.; Zhou, C.; Guo, Q.; Juefei-Xu, F.; Yu, H.; Feng, W.; Liu, Y.; and Wang, S. 2021. Auto-exposure fusion for single-image shadow removal. In *IEEE Conference on Computer Vision and Pattern Recognition*, 10571–10580.

Gryka, M.; Terry, M.; and Brostow, G. J. 2015. Learning to remove soft shadows. *ACM Transactions on Graph*, 34(5): 1–15.

Guo, L.; Huang, S.; Liu, D.; Cheng, H.; and Wen, B. 2023a. ShadowFormer: Global context helps shadow removal. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 710–718.

Guo, L.; Wang, C.; Yang, W.; Huang, S.; Wang, Y.; Pfister, H.; and Wen, B. 2023b. Shadowdiffusion: When degradation prior meets diffusion model for shadow removal. In *IEEE Conference on Computer Vision and Pattern Recognition*, 14049–14058.

Guo, L.; Zha, Z.; Ravishankar, S.; and Wen, B. 2021. Self-convolution: A highly-efficient operator for non-local image restoration. In *International Conference on Acoustics, Speech, and Signal Processing*, 1860–1864.

Guo, L.; Zha, Z.; Ravishankar, S.; and Wen, B. 2022. Exploiting non-local priors via self-convolution for highly-efficient image restoration. *IEEE Transactions on Image Processing*, 31: 1311–1324.

Guo, R.; Dai, Q.; and Hoiem, D. 2012. Paired regions for shadow detection and removal. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(12): 2956–2967.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Annual Conference on Neural Information Processing Systems*, 33: 6840–6851.

Hu, X.; Fu, C.-W.; Zhu, L.; Qin, J.; and Heng, P.-A. 2019. Direction-aware spatial context features for shadow detection and removal. *IEEE TPAMI*, 42(11): 2795–2808.

Jiang, H.; Luo, A.; Fan, H.; Han, S.; and Liu, S. 2023. Low-light image enhancement with wavelet-based diffusion models. *ACM Transactions on Graph*, 42(6): 1–14.

Jin, Y.; Sharma, A.; and Tan, R. T. 2021. Dc-shadownet: Single-image hard and soft shadow removal using unsupervised domain-classifier guided network. In *International Conference on Computer Vision*, 5027–5036.

Kingma, D. P.; and Ba, J. 2015. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*.

Le, H.; and Samaras, D. 2019. Shadow removal via shadow image decomposition. In *International Conference on Computer Vision*, 8578–8587.

Le, H.; and Samaras, D. 2020. From shadow segmentation to shadow removal. In *European Conference on Computer Vision*, 264–281.

Liu, F.; and Gleicher, M. 2008. Texture-consistent shadow removal. In *European Conference on Computer Vision*, 437–450. Springer.

Mohan, A.; Tumblin, J.; and Choudhury, P. 2007. Editing soft shadows in a digital photograph. *IEEE Computer Graphics and Applications*, 27(2): 23–31.

Nichol, A. Q.; and Dhariwal, P. 2021. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, 8162–8171.

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*.

Qu, L.; Tian, J.; He, S.; Tang, Y.; and Lau, R. W. 2017. Deshadownet: A multi-context embedding deep network for shadow removal. In *IEEE Conference on Computer Vision and Pattern Recognition*, 4067–4075.

Saharia, C.; Chan, W.; Chang, H.; Lee, C.; Ho, J.; Salimans, T.; Fleet, D.; and Norouzi, M. 2022a. Palette: Image-to-image diffusion models. In *ACM Transactions on Graph*, 1–10.

Saharia, C.; Ho, J.; Chan, W.; Salimans, T.; Fleet, D. J.; and Norouzi, M. 2022b. Image super-resolution via iterative refinement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4): 4713–4726.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Annual Conference on Neural Information Processing Systems*, 30.

Wan, J.; Yin, H.; Wu, Z.; Wu, X.; Liu, Y.; and Wang, S. 2022. Style-guided shadow removal. In *European Conference on Computer Vision*, 361–378. Springer.

Wang, J.; Li, X.; and Yang, J. 2018. Stacked conditional generative adversarial networks for jointly learning shadow detection and shadow removal. In *IEEE Conference on Computer Vision and Pattern Recognition*, 1788–1797.

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.

Xiao, C.; She, R.; Xiao, D.; and Ma, K.-L. 2013. Fast shadow removal using adaptive multi-scale illumination transfer. 32(8): 207–218.

Zhang, L.; Long, C.; Zhang, X.; and Xiao, C. 2020. Ris-gan: Explore residual and illumination with generative adversarial networks for shadow removal. In *AAAI Conference on Artificial Intelligence*, 12829–12836.

Zhang, L.; Zhang, Q.; and Xiao, C. 2015. Shadow remover: Image shadow removal based on illumination recovering optimization. *IEEE Transactions on Image Processing*, 24(11): 4623–4636.

Zhu, Y.; Huang, J.; Fu, X.; Zhao, F.; Sun, Q.; and Zha, Z.-J. 2022a. Bijective mapping network for shadow removal. In *IEEE Conference on Computer Vision and Pattern Recognition*, 5627–5636.

Zhu, Y.; Xiao, Z.; Fang, Y.; Fu, X.; Xiong, Z.; and Zha, Z.-J. 2022b. Efficient model-driven network for shadow removal. In *AAAI Conference on Artificial Intelligence*, 3635–3643.