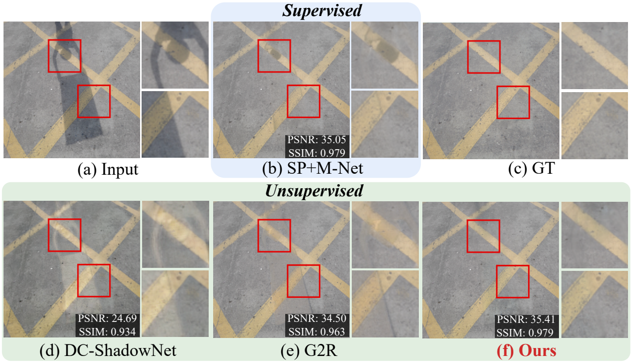
This ICCV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Boundary-Aware Divide and Conquer: A Diffusion-based Solution for Unsupervised Shadow Removal

Lanqing Guo[1], Chong Wang[1], Wenhan Yang[2], Yufei Wang[1], Bihan Wen[1]*
[1]Nanyang Technological University, Singapore [2]Peng Cheng Laboratory, China
{lanqing001, wang1711, yufei001, bihan.wen}@ntu.edu.sg, yangwh@pcl.ac.cn

Abstract

Recent deep learning methods have achieved superior results in shadow removal. However, most of these supervised methods rely on training over a huge amount of shadow and shadow-free image pairs, which require laborious annotations and may end up with poor model generalization. Shadows, in fact, only form partial degradation in images, while their non-shadow regions provide rich structural information potentially for unsupervised learning. In this paper, we propose a novel diffusion-based solution for unsupervised shadow removal, which separately modeling the shadow, non-shadow, and their boundary regions. We employ a pretrained unconditional diffusion model fused with non-corrupted information to generate the natural shadow-free image. While the diffusion model can restore the clear structure in the boundary region by utilizing its adjacent non-corrupted contextual information, it fails to address the inner shadow area due to the isolation of the non-corrupted contexts. Thus we further propose a Shadow-Invariant Intrinsic Decomposition module to exploit the underlying reflectance in the shadow region to maintain structural consistency during the diffusive sampling. Extensive experiments on the publicly available shadow removal datasets show that the proposed method achieves a significant improvement compared to existing unsupervised methods, and even is comparable with some existing supervised methods.



(a) Input (b) SP+M-Net (c) GT
(d) DC-ShadowNet (e) G2R (f) Ours

Figure 1: Shadow removal results on the ISTD [37] dataset. From (a) to (f): (a) input shadow image, supervised learning results of (b) SP+M-Net [25], (c) corresponding ground truth shadow-free image, as well as unsupervised learning results of (d) DC-ShadowNet [20], (e) G2R [29], and (f) Ours, respectively.

1. Introduction

Shadow is a ubiquitous phenomenon resulting from partial occlusion of light by occluders. It is critical to remove these shadows because their detrimental impacts on vision models, such as object detection and tracking [23, 35, 46]. Unfortunately, in general, it is still an open problem to remove shadows from a single image due to the large variety of shadow shapes and background structures, making it challenging to obtain a generalized solution. Conventional shadow removal methods [9, 45] mainly rely on carefully designed hand-crafted statistical features, *e.g.*, illumination, gradient, and region consistency, to construct the optimization function for shadow region removal. However, they totally ignore the natural image prior and the optimization function's underlying assumptions are frequently excessively idealistic, leading to unnatural results with artifacts, particularly in real-world scenarios.

Recently, deep-learning based image shadow removal methods [12, 25, 4, 7, 16, 28, 20] have achieved remarkable progress by learning the pixel-wise mapping between the shadow images and ground-truth shadow-free ones in a fully-supervised manner. However, such solutions require laborious annotations and can easily result in overfitting the training dataset with poor generalization. More importantly, shadow removal is a region-wise corrupted problem with abundant context and structure priors. This information actually provides rich clues to infer the shadow regions solely based on the single input, demonstrating strong potential to address the problem using unsupervised learning meth-

*Corresponding author: Bihan Wen.

ods. Some works [16, 28, 20] have started to explore unsupervised methods for shadow removal mainly relying on GAN using unpaired shadow and shadow-free images. Unfortunately, due to the absence of the training pair with the pixel-wise ground truth, the discriminator relies solely on unpaired non-shadow images, which can cause the generator to produce inauthentic outputs. Namely, the generator's learning faces dispersed space and is very easy to hallucinate new content and artifacts.

In this paper, we propose a novel unsupervised diffusion-based solution using only the shadow images without any reference. According to our analysis, the corrupted regions in shadow images can be categorized into two distinct types: 1) *shadow regions*, with subtle structural information obscured by the low illumination; 2) *boundary regions*, which contain noisy structures and exhibit rich adjacent non-corrupted contexts. Our work unifies the restoration of both shadow and boundary regions in a comprehensive manner, by integrating the generative power of the diffusion model and the detail preservation power of the intrinsic decomposition, resulting in mutual benefits for both regions.

More in detail, we employ a pre-trained unconditional diffusion model, injected with the guidance of the non-corrupted region information as the baseline to generate natural shadow-free images and suppress artifacts. While the diffusion model can effectively restore the clear structure and standard illumination in the boundary region by utilizing its adjacent non-corrupted contextual information, it falls short in addressing the inner shadow region, which is isolated from such contexts. To address this limitation, we propose a Shadow-Invariant Intrinsic Decomposition model, which ensures consistency among the reflectance of all intermediate results during diffusive sampling. By doing so, we are able to unveil the structural detail present in these inner shadow regions. Experimental results reveal that the proposed method consistently attains superior performance across existing widely-used shadow removal datasets. It markedly surpasses the capabilities of existing unsupervised methods, and in certain instances, achieves comparable performance to certain supervised methods.

The main contributions of this work are as follows:

- We propose a novel diffusion-based unsupervised method for shadow removal, in which we divide the corrupted regions into shadow and boundary regions. Inspired by the partition category, our work unifies shadow and boundary region restoration by integrating diffusion and intrinsic decomposition for their mutual benefits.

- We further present a Shadow-Invariant Intrinsic Decomposition model that guarantees coherence among reflectance values at each stage of diffusive sampling. This approach allows us to effectively uncover structural details present within the inner shadow regions.

- We conduct extensive experiments on public datasets and show that the proposed method achieves significant improvement among existing SOTA unsupervised methods and even comparable performance with some supervised methods.

2. Related Work

Shadow removal. Classic shadow removal methods rely on a series of prior information, such as gradient [9], illumination [45], and region consistency [12]. These methods are constructed based on the assumption of ideal conditions, which results in obvious shadow boundary artifacts when applied to real-world scenarios. Recent deep learning-based shadow removal methods [25, 6, 16, 27, 10, 18, 22, 11] boost the removal performance by relying on large-scale datasets of paired shadow and shadow-free images. However, as mentioned above, label annotations are difficult in practice and the model might be overfitted to the training set. More recently, inspired by unsupervised or weakly-supervised image translation methods [21, 19], [37, 16, 29] employ the GAN to generate shadow-free images with unpaired shadow and shadow-free images. For instance, Hu *et al.* [16] propose the mask-guided cycle-consistency constraint to simultaneously learn to produce shadows and remove shadows. Jin *et al.* [20] proposes the DC-ShadowNet to handle the soft and hard shadow removal using an unsupervised domain classifier. Specifically, Liu *et al.* [28] suggests discarding unpaired data in favor of employing a collection of shadow images. In this approach, the shadow generation sub-network transforms non-shadow regions into shadow ones, resulting in paired data suitable for training the shadow-removal sub-network. Nevertheless, the outcomes of these techniques consistently encountered issues with color distortions and the potential hallucination of new content and artifacts.

Denoising diffusion probabilistic models. Generative models have been widely applied to some low-level vision tasks, such as image super-resolution [34], inpainting [30], low-light enhancement [38] and deblurring [43] through conditional image generation. Very recently, diffusion probabilistic models, by modeling the conversion from a standard normal distribution to a data distribution $q(x)$ via diffusion process, have demonstrated impressive performance on generating high-quality images [13, 5, 14, 2, 33, 40]. However, training a task-specific conditional diffusion model from scratch demands substantial computational resources and can be time-consuming. Instead, a different line of research is to guide the sampling process of a pre-trained unconditional diffusion model to generate images with the desired semantics. Choi *et al.* [2] control the sampling process of a trained diffusion model using the low-frequency components of the reference images to generate the corresponding high-quality images. Wang *et al.* [39] adopt a range-null space decomposition to maintain the data consistency during the sampling for various image restoration tasks where only the null space contents are iteratively refined in the reverse diffusion process. Chung *et al.* [3] introduce a manifold constraint term to guide the sample path by alternating projection onto the measurement subspace, which demonstrated superior performance on various inverse problems. Most existing works focus on the restoration task with a specific degradation. In this paper, we explore the diffusion for shadow removal with uncertain degradation.

3. Preliminary

In this paper, we follow the diffusion model defined in [13]. The basic idea is to iteratively perturb a clean data sample $x_0 \sim q(x)$ with small Gaussian noise in T steps, producing a sequence of noise step $\{x_t\}_{t=1}^{T}$ with the corresponding noise scale step $\{\beta_t\}_{t=1}^{T}$ during the forward process, which can be described as Gaussian transition:

$$q(x_t|x_{t-1}) = \mathcal{N}\left(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t\mathbf{I}\right). \tag{1}$$

A nice property of the forward process is that the noisy data x_t can be sampled from x_0 in a closed form using reparameterization:

$$q(x_t|x_0) = \mathcal{N}\left(x_t; \sqrt{\bar{\alpha}_t}x_0, (1-\bar{\alpha}_t)\,\mathbf{I}\right), \tag{2}$$

where $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i$.

The diffusion model is trained to approximate the reverse process of (1) from a pure Gaussian noise $x_T \sim \mathcal{N}(0, \mathbf{I})$ to a clean sample x_0. The estimation of the previous state x_{t-1} can be derived from the posterior distribution $p(x_{t-1}|x_t, x_0)$ as follows:

$$p(x_{t-1}|x_t, x_0) = \mathcal{N}\left(x_{t-1}; \mu_t(x_t, x_0), \sigma^2\mathbf{I}\right). \tag{3}$$

Specifically, a noise predictor ϵ_θ is trained to estimate the parameters $\mu_\theta(x_t, t)$ and $\Sigma_\theta(x_t, t)$ of the reverse Gaussian distribution at time step t,

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}\left(x_{t-1}; \mu_\theta\left(x_t, t\right), \Sigma_\theta\left(x_t, t\right)\right), \tag{4}$$

where the predicted mean $\mu_\theta(x_t, t)$ can be parameterized by the noise predictor ϵ_θ as: $\mu_\theta(x_t, t) = \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(x_t, t)\right)$. Ho *et al.* [13] set a fixed variance $\Sigma_\theta(x_t, t) = \beta_t^2$ to simplify the training objective while Nichol *et al.* [31] adopt a learned variance $\Sigma_\theta(x_t, t)$ in the reverse process (4) to reduce the number of sampling steps.

4. Methodology

4.1. Motivation

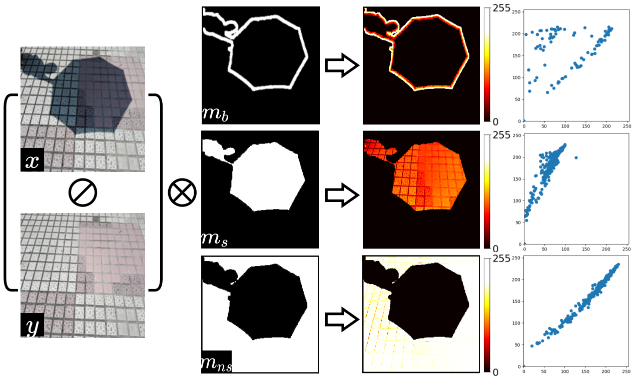


Figure 2: Analysis of the illumination transformation from shadow image x to ground truth shadow-free image y for mask-divided three regions, *i.e.*, (1) the boundary regions cropped by m_b, (2) the shadow regions cropped by m_s, (3) the non-shadow regions cropped by m_{ns}. $\oslash$ and $\otimes$ denote the element-wise division and multiplication, respectively. We randomly sample 1000 pixels for each region and illustrate the illumination mapping in the last column. It indicates the illumination transformation in shadow and non-shadow regions can be approximated by a simple linear transformation, while the illumination transformation in boundary regions is sophisticated and difficult to model. Inspired by this observation, our work adopts a boundary-aware divide and conquer methodology to deal with shadow and boundary regions separately.

Here we illustrate the motivation behind the design of our shadow removal algorithm. We denote the shadow image as x, the corresponding shadow mask is m and the shadow-free image is y. We define a boundary extractor $B(\cdot)$ to extract the penumbra (boundary) area as the residual of dilated and eroded mask, achieving the boundary mask $m^b = B(m)$. The shadow image can be divided into three regions: shadow (umbra) region x^s, boundary (penumbra) region x^b, and non-shadow region x^{ns} as shown in Figure 2. Shadow removal is a restoration problem that faces a region-dependent corruption problem. It can be addressed via a conditionally region-based inpainting problem. *First, for different regions, their utilized information for shadow removal is varied.* While the contextual non-shadow regions and underlying structural information in shadow regions improve predictions of corrupted regions (*i.e.*, x^{b+s}), the contribution of information from different areas varies in shadow removal. The boundary regions have lower prediction uncertainty due to rich adjacent information, while inner shadow regions have higher uncertainty being far from valid regions. *That is to say, more information and constraints should be introduced to facilitate the restoration of inner shadow regions. Second, the Retinex model can well disentangle the shadow illumination change while maintaining the intrinsic structure constantly.* Based on the Retinex theory [24], an input image can be decomposed into a product of a reflectance image and an illumination image. As shown in Figure 2, the shadow-to-shadow-free

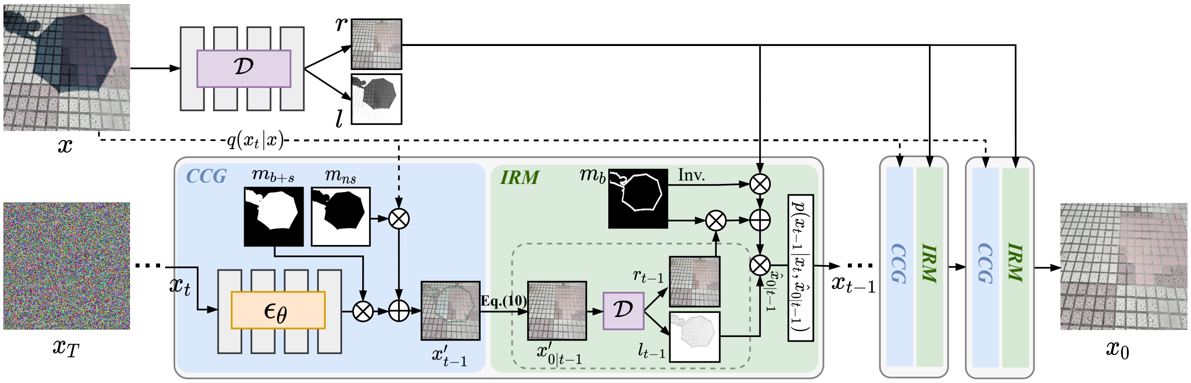


Figure 3: Overview of our Boundary-aware Conditional Diffusion (BCDiff) approach. For state x_t, BCDiff consists of two steps: (1) CCG: context conditioned generation: we first simultaneously sample the non-corrupted region and corrupted region according to Eq. (5) and Eq. (6), respectively, achieving the intermediate image x'_{t-1}; (2) IRM: iterative reflectance maintenance: we then maintain the structural information consistency by boundary-aware weighted integrating the decomposed reflectance of $x'_{0|t-1}$ ('clean' version of x'_{t-1}) and the reflectance of the original shadow image x.

image illumination transformation is a simple linear process in shadow regions, but modeling the sophisticated and sharp changes around the shadow boundary in boundary regions is difficult. Meanwhile, ideally, the shadow degradation only affects the illumination map while the reflectance map is constantly across different shadows. *In other words, the reflectance map provides a reliable clue to infer the structural details in the sampling dynamics.*

4.2. Boundary-Aware Conditional Diffusion

We propose unsupervised **B**oundary-aware **C**onditional **Diff**usion (BCDiff) method for shadow removal is illustrated in Figure 3 and summarized in Algorithm 1. The shadow removal can be re-formulated as a conditionally region-based inpainting, where the restoration of both shadow and boundary regions is unified in a mutually beneficial manner, by integrating a pre-trained unconditional denoising diffusion probabilistic model (4) and the intrinsic decomposition, introducing more information and constraints to recover inner shadow regions.

Context conditioned generation. Intuitively, the non-shadow region is non-corrupted and we can sample the intermediate image x_t^{ns} at any timestep t on using (2) as follows:

$$x_{t-1}^{\text{ns}} \sim \mathcal{N}\left(\sqrt{\bar{\alpha}_t}x, (1-\bar{\alpha}_t)\mathbf{I}\right) \ . \tag{5}$$

While the corrupted regions, *i.e.*, boundary and shadow regions denoted as x^{b+s}, can be sampled using (4) as follow:

$$x_{t-1}^{\text{b+s}} \sim \mathcal{N}\left(\mu_\theta\left(x_t, t\right), \Sigma_\theta\left(x_t, t\right)\right) \ . \tag{6}$$

To this end, we obtain the whole intermediate image x'_{t-1} by spatially combining the corrupted and non-corrupted regions via mask:

$$x'_{t-1} = m^{b+s} \circ x_{t-1}^{\text{b+s}} + (1 - m^{b+s}) \circ x_{t-1}^{\text{ns}} \ , \tag{7}$$

where the m^{b+s} indicates the corrupted regions.

Illumination-consistency constraint. We constrain the illumination consistency between shadow and non-shadow regions during the process of diffusion sampling. To pursue illumination consistency, we calculate the mean value of shadow and non-shadow regions in x_t to approximate their illumination. Inspired by [5], where the gradient of a classifier is used for conditioning the diffusion generation, we incorporate the gradient of the loss measuring the difference between the mean value of the shadow and non-shadow regions and extend (6) as follows:

$$\hat{\epsilon} = \epsilon_\theta(x_t, t) - \sqrt{1-\bar{\alpha}_t}\nabla_{x_t}|u_t^s - u_t^{ns}| \ , \tag{8}$$

$$x_{t-1}^{b+s} = \frac{1}{\sqrt{\bar{\alpha}_t}}\left(x_t - \frac{\beta_t}{\sqrt{\bar{\alpha}_{t-1}}}\hat{\epsilon}\right) + \sigma_t z, \quad z \sim \mathcal{N}\left(\mathbf{0}, \mathbf{I}\right), \tag{9}$$

where u_t^s and u_t^{ns} calculate the mean value of the shadow region and non-shadow region in x_t, respectively. By utilizing the gradient during the sampling process, we can thus control the generation results of the pre-trained diffusion model to have coherent illumination.

Iterative reflectance maintenance. Another key point for shadow removal is preserving the structural information in shadow images during generation. Different from the boundary regions that have rich adjacent non-corrupted information, the inner shadow regions are always isolated from the non-corrupted regions. Thus, we exploit the structural information hidden in shadow regions as the auxiliary to further constrain the fidelity. Here we introduce a

Algorithm 1 Boundary-aware conditional diffusion.

Input: shadow image x, shadow mask m, pre-trained unconditional diffusion model ϵ_θ, pre-trained decomposition model $\mathcal{D}$, number of implicit sampling iterations T.

```
$x_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
$r, l = \mathcal{D}(x)$
for $t = T, \dots, 1$ do
    $x_{t-1}^{\text{ns}} \sim \mathcal{N}\left(\sqrt{\bar{\alpha}_t} x, (1 - \bar{\alpha}_t)\, \mathbf{I}\right)$
    $\hat{\epsilon} = \epsilon_\theta(x_t, t) - \sqrt{1-\bar{\alpha}_t} \nabla_{x_t} |u_t^s - u_t^{ns}|$
    $z \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ if $t > 1$, else $z = \mathbf{0}$
    $x_{t-1}^{b+s} = \frac{1}{\sqrt{\bar{\alpha}_t}} \left( x_t - \frac{\beta_t}{\sqrt{\bar{\alpha}_{t-1}}} \hat{\epsilon} \right) + \sigma_t z$
    $x'_{t-1} = m \circ x_{t-1}^{\text{b+s}} + (1 - m) \circ x_{t-1}^{\text{ns}}$
    $x'_{0|t-1} = \frac{1}{\sqrt{\bar{\alpha}_t}} \left( x'_{t-1} - \hat{\epsilon} \sqrt{1 - \bar{\alpha}_t} \right)$
    $r_{t-1}, l_{t-1} = \mathcal{D}(x'_{0|t-1})$
    $\hat{x}_{0|t-1} = m_b \circ r_{t-1} \circ l_{t-1} + (1 - m_b) \circ r \circ l_{t-1}$
    $x_{t-1} \sim p(x_{t-1} | x_t, \hat{x}_{0|t-1})$
end for
return $x_0$
```

Shadow-Invariant Intrinsic Decomposition (SIID) model $\mathcal{D}$ (details refer to Sec. 4.3) to decompose the reflectance and illumination maps $\{r, l\}$ and $\{r_{t-1}, l_{t-1}\}$ of the original shadow image x and the intermediate image x'_{t-1}, respectively. To mitigate the impact of the noise in x_t, we first estimate its intermediate clean image $x'_{0|t-1}$ by reversing the process (2). The whole process of iterative reflectance maintenance can be formulated as :

$$x'_{0|t-1} = \frac{1}{\sqrt{\bar{\alpha}_t}} \left(x'_{t-1} - \hat{\epsilon} \sqrt{1 - \bar{\alpha}_t} \right) \ , \tag{10}$$

$$r, l = \mathcal{D}(x) \quad r_{t-1}, l_{t-1} = \mathcal{D}(x'_{0|t-1}) \ . \tag{11}$$

During the iterative process of diffusion sampling, the reflectance of x_{t-1} should be consistent with the original shadow image x within shadow regions since the shadow ideally only corrupts the illumination map. We spatially integrate the decomposed reflectance map r_{t-1} of each timestep and the original r of the shadow image according to boundary mask m_b to separately restore the boundary and shadow regions:

$$\hat{x}_{0|t-1} = m_b \circ r_{t-1} \circ l_{t-1} + (1 - m_b) \circ r \circ l_{t-1} \ . \tag{12}$$

Then we yield x_{t-1} by sampling from $p(x_{t-1} | x_t, \hat{x}_{0|t-1})$.

4.3. Shadow-Invariant Intrinsic Decomposition

Intrinsic image decomposition [24, 1] factorizes an input image v into a product of a reflectance image and an illumination image: $v = r \circ l$. A shadow-invariant intrinsic decomposition (SIID) model is introduced for unveiling structures of inner shadow regions in the diffusion sampling process, which is illustrated in Figure 4.

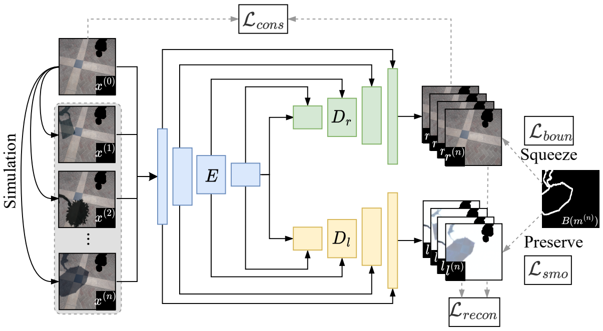


Figure 4: Overview of the shadow-invariant intrinsic decomposition model $\mathcal{D}$. During the training stage, the input is the simulated image set $\{x^{(i)}\}_{i=0}^n$ from shadow image x_0. $\mathcal{D}$ includes a decoder E, as well as two decoders D_r and D_l for the reflectance $\{r^{(i)}\}_{i=0}^n$ and illumination $\{l^{(i)}\}_{i=0}^n$.

Shadow image set simulation. Thus, we propose to synthesize shadows in the non-shadow regions to construct the image set $\mathcal{X}$ with the same scene but various shadow conditions. In detail, given a shadow image x and corresponding shadow mask m, we simulate n different shadows within the non-shadow background, denoting as $x^{(0)} = (1-m) \circ x$ as shown in Figure 4. The shadow synthesis can be formulated as $x^{(i)} = \phi\left(x^{(0)}, m^{(i)}, \theta^{(i)}\right)$ with $i \in \{1, 2, \dots, n\}$, where $\phi(\cdot)$ denotes the shadow synthetic algorithm [17]. $\theta^{(i)}$ denotes the random pre-defined parameter to simulate different shadows and $m^{(i)}$ denotes the indexed binarized patterns from the external mask set $\mathcal{M}$ for simulation. After that, we train the SIID model with the synthesized image set $\{x^{(i)}\}_{i=0}^n$. According to the SIID model, we obtain the decomposed reflectance set $\{r^{(i)}\}_{i=0}^n$ and illumination set $\{l^{(i)}\}_{i=0}^n$ for simulated image set. Details of the decomposition architectures and shadow simulation process are provided in **supplementary**.

Loss functions. Since reflectance is constant for different shadow conditions, we should be able to use the reflectance $r^{(i)}$ predicted by *any* image $x^{(i)} \in \mathcal{X}$ to reconstruct $x^{(j)}$, when paired with $l^{(j)}$, as following:

$$\mathcal{L}_{recon} = \sum_{i=1}^{n} \sum_{j=1}^{n} \left\| r^{(i)} \circ l^{(j)} - x^{(j)} \right\|_1 . \tag{13}$$

Besides, we regard the non-shadow background with normal illumination as the ground truth reflectance map. Thus, we also include a reflectance consistency loss that constrains the predicted reflectances should be identical:

$$\mathcal{L}_{cons} = \sum_{i=1}^{n} \left\| r^{(i)} - r^{(0)} \right\|_1 \ . \tag{14}$$

The illumination map should be locally consistent for the surface of each object in the scene [1], thus we utilize the total variation minimization (TV) to minimize the gradient of the predicted illumination map excluded the boundary area. In the meanwhile, we also adopt a boundary smoothness loss to 'squeezing' boundary trace from the reflectance map as following

$$\mathcal{L}_{smo} = \sum_{i=1}^{n} \left\| \nabla l^{(i)} \circ \left(1 - B(m^{(i)})\right) \right\|_1 , \tag{15}$$

$$\mathcal{L}_{boun} = \sum_{i=1}^{n} \left\| \nabla r^{(i)} \circ B(m^{(i)}) \right\|_1 , \tag{16}$$

where ∇ stands for the gradient including horizontal ∇_h and vertical ∇_v.

The hybrid objective function $\mathcal{L}_{total}$ is obtained by combining the above losses, which guides the training of the decomposition model $\mathcal{D}$ as follows,

$$\mathcal{L}_{total} = \mathcal{L}_{recon} + \lambda_1 \mathcal{L}_{cons} + \lambda_2 \mathcal{L}_{smo} + \lambda_3 \mathcal{L}_{boun} , \tag{17}$$

where λ_1, λ_2, and λ_3 are the weighting coefficients to balance the influence of each term.

5. Experiments

5.1. Datasets and Evaluation Metrics

Datasets. The ISTD dataset includes 1,870 triplets of shadow, shadow mask, and shadow-free images with the solution of 480×640, where 1,330 triplets are for training and 540 for testing. Since the unsupervised method does not learn the standard illumination from real shadow-free samples, we follow [25, 29] to apply the adjusted testing set with reduced illumination difference within non-shadow regions between the shadow and shadow-free images in the original dataset.

Evaluation Metrics. Following the previous works [37, 12, 32, 25, 4, 7], we use the Root-Mean-Square Error (RMSE), Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity (SSIM) [41] as the evaluation metrics. We compute the RMSE in the LAB color space, and PSNR and SSIM scores in the RGB color space to evaluate our method. The lower RMSE values indicate less reconstruction error between the recovered output and ground truth, while higher values indicate better results for the PSNR and SSIM.

5.2. Experimental Settings

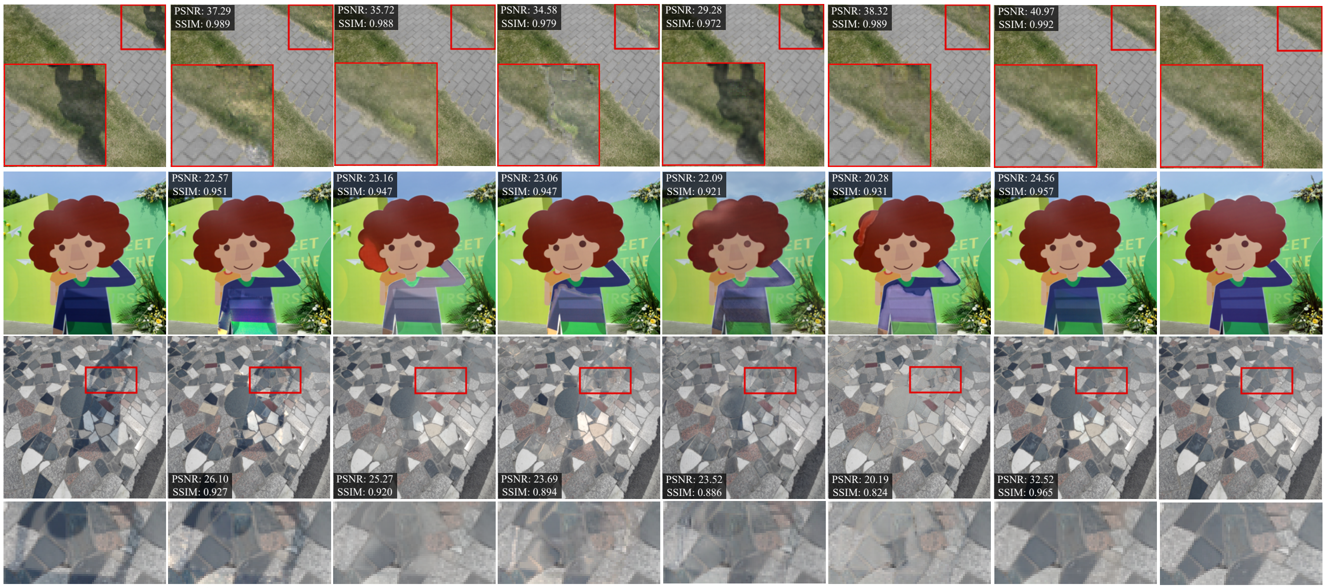
The proposed method is implemented using PyTorch, which is conducted on one NVIDIA RTX A5000 GPU. Our shadow removal method is training-free and relies on a pre-trained unconditional diffusion model [13]. We use $T = 250$ timesteps for all our experiments. The sampling batch size is set to 3. The kernels of both dilation and erosion operations in $B(\cdot)$ are disks with a radius of 8 pixels. The decomposition model is modified on the most recent transformer-based image-to-image backbone [42], whose detailed architectures are provided in **supplementary**. The weights of our losses for training decomposition model $\{\lambda_1, \lambda_2, \lambda_3\}$ are set empirically to $\{1, 1, 0.1\}$. We set the number of simulated shadow images as $n = 4$.

5.3. Comparison with the-state-of-the-arts

We compare our proposed unsupervised method with several state-of-the-art methods, including classic methods: Gong *et al.* [8], Yang *et al.* [44], and Guo *et al.* [12]; supervised methods (training with paired shadow and shadow-free images): DSC [15], DHAN [4], SP+M-Net [25], Fu *et al.* [7], BMNet [49], and SG-ShadowNet [36]; unsupervised methods (training without paired shadow and shadow-free images): MaskShadow-GAN [16], LG-ShadowNet [28], DC-ShadowNet [20], Le *et al.* [26], and G2R [29]. All of the shadow removal results by the competing methods are quoted from the original papers or reproduced using their official implementations. We evaluate the performance with a resolution of 256×256 following most previous methods [7, 25, 29].

Table 1 shows the quantitative results of the testing set of the ISTD dataset. Gong *et al.* [8] also specify the task by a series of pre-defined priors, which is too strict and hard to extend to real-world scenarios. Those supervised methods share the same type of training data, including shadow and shadow-free image pairs. They learn the mapping from shadow image to shadow-free one according to the training pairs. However, their performance might be largely degraded when extended to some unseen scenes. With the merits of exploiting contextual non-shadow regions and adapting the decomposition model to the testing data, our proposed unsupervised method without GT in the training stage can even achieve better results compared to some supervised methods. Besides, some unsupervised methods, *e.g.*, MaskShadow-GAN [16], LG-ShadowNet [28], and DC-ShadowNet [20], train their shadow removal models using unpaired shadow and shadow-free images. We can see that our method outperforms these three methods although learning only with shadow images. The setting of Le *et al.* [26] and G2R [29] is the same as ours, which is fairer since they all do not require any shadow-free images. Compared to the state-of-the-art unsupervised methods G2R [29], our results for the shadow regions are better by around 2.5dB in PSNR. Besides, following the previous method [49], we also verify our performance when the network takes the imperfect detected shadow masks from [47], denoted as 'w/ detected mask' in Table 1. With the input of the detected masks, the performance of our method will be slightly degraded.

Figure 5 shows the qualitative results of our method and the other state-of-the-art methods on the ISTD dataset. For some examples with small shadow regions as shown in the



(a) Input (b) Gong *et al.* (c) SP+M-Net (d) Le *et al.* (e) DC-ShadowNet (f) G2R (g) Ours (h) GT

Figure 5: Shadow removal results on ISTD [37] dataset. From (a) to (h): (a) input shadow image, and the result of classic method (b) Gong *et al.* [8]; supervised learning result of (c) SP+M-Net [25]; unsupervised learning results of (d) Le *et al.* [26], (e) DC-ShadowNet [20], (f) G2R [29], and (g) Ours, as well as (h) ground truth shadow-free image, respectively.

Method	Setting	Shadow Region (S)			Non-Shadow Region (NS)			All Image (ALL)		
		PSNR↑	SSIM↑	RMSE↓	PSNR↑	SSIM↑	RMSE↓	PSNR↑	SSIM↑	RMSE↓
Yang *et al.* [44]	Classic	21.57	0.878	23.2	22.25	0.782	14.2	20.26	0.706	15.9
Gong *et al.* [8]		30.53	0.972	13.0	36.63	0.982	2.6	28.96	0.943	4.3
Guo *et al.* [12]		26.89	0.960	20.1	35.48	0.975	3.1	25.51	0.924	6.1
DHAN [4]	Supervised	32.92	0.988	9.6	27.15	0.971	7.4	25.66	0.956	7.8
SP+M-Net [25]		**37.60**	**0.990**	6.3	**36.02**	0.976	3.0	**32.94**	**0.962**	3.5
Fu *et al.* [7]		36.04	0.978	6.7	31.16	0.892	3.8	29.45	0.861	4.2
BMNet [49] (w/ detected mask)		-	-	6.1	-	-	2.9	-	-	3.5
BMNet [49] (w/ GT mask)		-	-	**5.6**	-	-	**2.5**	-	-	**3.0**
SG-ShadowNet [36]		36.80	**0.990**	6.5	35.57	**0.978**	2.9	32.46	**0.962**	3.4
MaskShadow-GAN [16]	Unsupervised	32.19	0.984	10.8	33.44	0.974	3.8	28.81	0.946	4.8
LG-ShadowNet [28]		32.44	0.982	9.9	33.68	0.971	3.4	29.20	0.945	4.4
DC-ShadowNet [20]		31.06	0.976	12.2	27.03	0.961	6.8	25.03	0.926	7.8
Le *et al.* [26]		33.09	0.983	10.4	35.26	0.977	2.9	30.12	0.950	4.0
G2R [29]		33.58	0.979	8.9	35.52	0.976	2.9	30.52	0.944	3.9
Ours (w/ detected mask)		35.71	**0.986**	**7.6**	36.39	0.981	2.7	32.11	0.959	3.5
Ours (w/ GT mask)		**35.91**	**0.986**	**7.6**	**37.27**	**0.984**	**2.4**	**32.73**	**0.962**	**3.3**

Table 1: Quantitative comparison results of the proposed method with the state-of-the-art methods on ISTD [37] dataset. The best and second performances for supervised learning and unsupervised learning methods are highlighted in **Bold** and underlined, respectively. '-' denotes the results are not publicly available.

first row in Figure 5, the unsupervised methods have better performance because of rich context information in shadow images. Our method can better preserve the structural information and suppress the boundary artifacts even compared with SP+M-Net (supervised method). Besides, the restored results of our method would be more natural than the competing methods as shown in the second row in Figure 5. The results of some existing methods, *e.g.*, SP+M-Net [25], Le *et al.* [26], and G2R [28], are brightened but not consistent with the surrounding colors and illumination, while our method can obtain results with better consistency.

5.4. Ablation Study

To demonstrate the effectiveness of each key component of the proposed method, we conduct experiments on several model variants on the ISTD dataset. We also provide the ablation studies for the effects of losses in the decomposition model in **supplementary**.

The effect of the diffusion model. We propose to utilize the diffusion model (sampling process) to act as a natural image prior to suppressing the artifacts of generated results and correcting the illumination of shadow regions according to the context information. We conduct experiments to verify the effect of the diffusion model on the ISTD dataset. In Table 2, we provide the shadow removal performance without the diffusion model, where we only preserve the decomposition model and regard the decomposed reflectance maps as the restored shadow-free results. We found that the performance has dropped on all metrics, especially for the SSIM metric from 0.962 to 0.952. In addition, we illustrate the visual comparisons without and with the diffusion model in Figure 6(e) and (h). We can see that there will be obvious boundary artifacts and illumination inconsistency between shadow and non-shadow regions without the diffusion model.

The effect of iterative reflectance maintenance. To generate the shadow-free image conditioned by the underlying structural information hidden in the shadow regions, we propose iterative reflectance maintenance during the diffusive sampling. We conduct experiments to remove the reflectance replacement for each iteration, where we dropped out the (7) during the sampling stage (denoted as w/o reflectance maintain in Table 2). We find a very obvious performance drop on all metrics, indicating that the generated shadow-free images lose huge fidelity as shown in Figure 6(f). It also verified that even without the guidance of structural information in corrupted regions, the illumination information would still be easier to predict according to the contextual non-corrupted information.

The effect of the simulated image set. We explore whether the simulated image set for one scene can improve the decomposition performance compared to using one simulated shadow image. Specifically, we set the number of simulated shadow images as $n = 1$ (denoted as w/o simulated set in Table 2), the performance will decrease especially for shadow regions. Without learning the various illumination conditions in the training stage, the decomposition model will easily wrongly exclude some structural and color components from reflectance due to the domain gap between simulated and real shadow images, leading to serious artifacts in results as shown in Figure 6 (f).

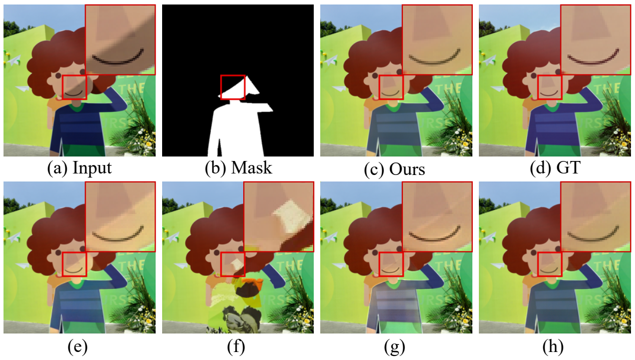
The effect of the illumination-consistency constraint. We also conduct experiments to verify the effectiveness of the illumination consistency constraint during the diffusive sampling process (denoted as w/o illumination constraint in Table 2). In details, we replace the Eqs. (8) and (9) with original denoising step in DDPM:

$$x^s_{t-1} = \frac{1}{\sqrt{\bar{\alpha}_t}}\left(x_t - \frac{\beta_t}{\sqrt{\bar{\alpha}_{t-1}}}\epsilon_\theta(x_t,t)\right) + \sigma_t z, \qquad (18)$$

where $z \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. The performance would be degraded without the illumination constraint, indicating that the illumination constraint can further enforce the illumination consistency of the restored results.

Method	Shadow		All	
	PSNR↑	SSIM↑	PSNR↑	SSIM↑
w/o simulated set	35.16	0.982	32.13	0.953
w/o diffusion model	34.87	0.980	31.89	0.952
w/o reflectance maintain	27.25	0.910	26.01	0.871
w/o illumination constraint	35.85	0.984	32.40	0.959
Complete model	**35.91**	**0.986**	**32.73**	**0.962**

Table 2: Ablation study to verify the effectiveness of each component in our method.



(a) Input (b) Mask (c) Ours (d) GT

(e) (f) (g) (h)

Figure 6: Visual comparisons of (a) input shadow image, (b) corresponding shadow mask, (c) the result of our complete model, (d) ground truth shadow-free image, as well as the results of our model variants (e) w/o diffusion model, (f) w/o reflectance maintain, (g) w/o simulated image set, and (h) w/o illumination constraint.

Method	RMSE↓	PSNR↑	SSIM↑
SP+M-Net [25]	22.2	-	-
Le *et al.* [26]	20.9	-	-
Mask-ShadowGAN [16]	19.6	20.38	0.887
LG-ShadowNet [28]	18.3	20.68	0.880
G2R [29]	18.8	21.07	0.882
Ours	**17.7**	**22.23**	**0.893**

Table 3: The quantitative comparison of video shadow removal using our method and recent state-of-the-art unsupervised methods. Note that the metrics are only in the moving-shadow regions following previous methods [29]. '-' denotes the results are not publicly available.

5.5. Video Shadow Removal

We further evaluate our method of video shadow removal. We select the public available video shadow removal dataset [26], which contains 8 videos whose contents are static scenes with invariant backgrounds. This dataset also provides a corresponding V_{max} as the pseudo shadow-free frame for each video according to taking the maximum intensity values at each pixel location across the whole video. The mask for moving shadows encompasses pixels present in both the shadow and non-shadow areas of the video, delineating the evaluation region. We adopt the configuration outlined in the official code of the work by [26] using a threshold of 80 to generate the moving shadow mask.

For this data, we apply the pre-trained shadow detector [48] to generate the shadow masks for our experiments following most previous methods [28]. Table 3 summarizes the video shadow removal performance of our method and recent unsupervised methods, where our method outperforms all competing methods on all metrics.

6. Conclusion

In this paper, we present a novel diffusion-based unsupervised method for shadow removal. We employ a pre-trained unconditional diffusion fused with non-corrupted information as the baseline to generate natural shadow-free images. Based on that, we propose iterative reflectance maintenance under the auxiliary of the shadow-invariant intrinsic decomposition model to preserve the underlying structures within shadow regions, as well as the illumination consistency constraint to pursue consistent illumination across different regions. Finally, comprehensive experiments demonstrate the superiority of our method and the better generalizability to unseen scenes, which achieves significant improvement compared to the state-of-the-art unsupervised methods over publicly available datasets, and even is comparable with some existing supervised methods.

Acknowledgement

This work was supported in part by the Ministry of Education, Republic of Singapore, through its Start-Up Grant and Academic Research Fund Tier 1 (RG61/22), and in part by the Basic and Frontier Research Project of PCL, and the Major Key Project of PCL. This work was carried out at the ROSE Lab, Nanyang Technological University, Singapore.

References

[1] Yu-Sheng Chen, Yu-Ching Wang, Man-Hsin Kao, and Yung-Yu Chuang. Deep photo enhancer: Unpaired learning for image enhancement from photographs with gans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6306–6314, 2018. 5

[2] Jooyoung Choi, Sungwon Kim, Yonghyun Jeong, Youngjune Gwon, and Sungroh Yoon. Ilvr: Conditioning method for denoising diffusion probabilistic models. *arXiv preprint arXiv:2108.02938*, 2021. 2

[3] Hyungjin Chung, Byeongsu Sim, Dohoon Ryu, and Jong Chul Ye. Improving diffusion models for inverse problems using manifold constraints. *arXiv preprint arXiv:2206.00941*, 2022. 3

[4] Xiaodong Cun, Chi-Man Pun, and Cheng Shi. Towards ghost-free shadow removal via dual hierarchical aggregation network and shadow matting gan. In *AAAI*, pages 10680–10687, 2020. 1, 6, 7

[5] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021. 2, 4

[6] Bin Ding, Chengjiang Long, Ling Zhang, and Chunxia Xiao. Argan: Attentive recurrent generative adversarial network for shadow detection and removal. In *ICCV*, pages 10213–10222, 2019. 2

[7] Lan Fu, Changqing Zhou, Qing Guo, Felix Juefei-Xu, Hongkai Yu, Wei Feng, Yang Liu, and Song Wang. Auto-exposure fusion for single-image shadow removal. In *CVPR*, pages 10571–10580, 2021. 1, 6, 7

[8] Han Gong and Darren Cosker. Interactive removal and ground truth for difficult shadow scenes. *JOSA A*, 33(9):1798–1811, 2016. 6, 7

[9] Maciej Gryka, Michael Terry, and Gabriel J Brostow. Learning to remove soft shadows. *ACM Transactions on Graphics*, 34(5):1–15, 2015. 1, 2

[10] Lanqing Guo, Siyu Huang, Ding Liu, Hao Cheng, and Bihan Wen. Shadowformer: Global context helps image shadow removal. *arXiv preprint arXiv:2302.01650*, 2023. 2

[11] Lanqing Guo, Chong Wang, Wenhan Yang, Siyu Huang, Yufei Wang, Hanspeter Pfister, and Bihan Wen. Shadowdiffusion: When degradation prior meets diffusion model for shadow removal. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14049–14058, 2023. 2

[12] Ruiqi Guo, Qieyun Dai, and Derek Hoiem. Paired regions for shadow detection and removal. 35(12):2956–2967, 2012. 1, 2, 6, 7

[13] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 2, 3, 6

[14] Jonathan Ho, Chitwan Saharia, William Chan, David J Fleet, Mohammad Norouzi, and Tim Salimans. Cascaded diffusion models for high fidelity image generation. *J. Mach. Learn. Res.*, 23(47):1–33, 2022. 2

[15] Xiaowei Hu, Chi-Wing Fu, Lei Zhu, Jing Qin, and Pheng-Ann Heng. Direction-aware spatial context features for shadow detection and removal. 42(11):2795–2808, 2020. 6

[16] Xiaowei Hu, Yitong Jiang, Chi-Wing Fu, and Pheng-Ann Heng. Mask-shadowgan: Learning to remove shadows from unpaired data. In *ICCV*, pages 2472–2481, 2019. 1, 2, 6, 7, 8

[17] Naoto Inoue and Toshihiko Yamasaki. Learning from synthetic shadows for shadow detection and removal. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(11):4187–4197, 2020. 5

[18] Yeying Jin, Ruoteng Li, Wenhan Yang, and Robby T Tan. Estimating reflectance layer from a single image: Integrating reflectance guidance and shadow/specular aware learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 1069–1077, 2023. 2

[19] Yeying Jin, Beibei Lin, Wending Yan, Wei Ye, Yuan Yuan, and Robby T. Tan. Enhancing visibility in nighttime haze images using guided apsf and gradient adaptive convolution, 2023. 2

[20] Yeying Jin, Aashish Sharma, and Robby T Tan. Dc-shadownet: Single-image hard and soft shadow removal using unsupervised domain-classifier guided network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5027–5036, 2021. 1, 2, 6, 7

[21] Yeying Jin, Wenhan Yang, and Robby T Tan. Unsupervised night image enhancement: When layer decomposition meets light-effects suppression. In *European Conference on Computer Vision*, pages 404–421. Springer, 2022. 2

[22] Yeying Jin, Wenhan Yang, Wei Ye, Yuan Yuan, and Robby T Tan. Shadowdiffusion: Diffusion-based shadow removal using classifier-driven attention and structure preservation. *arXiv preprint arXiv:2211.08089*, 2022. 2

[23] Cláudio Rosito Jung. Efficient background subtraction and shadow removal for monochromatic video sequences. 11(3):571–577, 2009. 1

[24] Edwin H Land. The retinex theory of color vision. *Scientific american*, 237(6):108–129, 1977. 3, 5

[25] Hieu Le and Dimitris Samaras. Shadow removal via shadow image decomposition. In *ICCV*, pages 8578–8587, 2019. 1, 2, 6, 7, 8

[26] Hieu Le and Dimitris Samaras. From shadow segmentation to shadow removal. In *ECCV*, 2020. 6, 7, 8

[27] Hieu Le and Dimitris Samaras. Physics-based shadow image decomposition for shadow removal. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12):9088–9101, 2021. 2

[28] Zhihao Liu, Hui Yin, Yang Mi, Mengyang Pu, and Song Wang. Shadow removal by a lightness-guided network with training on unpaired data. *IEEE Transactions on Image Processing*, 30:1853–1865, 2021. 1, 2, 6, 7, 8, 9

[29] Zhihao Liu, Hui Yin, Xinyi Wu, Zhenyao Wu, Yang Mi, and Song Wang. From shadow generation to shadow removal. In *CVPR*, 2021. 1, 2, 6, 7, 8

[30] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11461–11471, 2022. 2

[31] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, pages 8162–8171. PMLR, 2021. 3

[32] Liangqiong Qu, Jiandong Tian, Shengfeng He, Yandong Tang, and Rynson WH Lau. Deshadownet: A multi-context embedding deep network for shadow removal. In *CVPR*, pages 4067–4075, 2017. 6

[33] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 2

[34] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 2

[35] Andres Sanin, Conrad Sanderson, and Brian C Lovell. Improved shadow removal for robust person tracking in surveillance scenarios. In *International Conference on Pattern Recognition*, pages 141–144. IEEE, 2010. 1

[36] Jin Wan, Hui Yin, Zhenyao Wu, Xinyi Wu, Yanting Liu, and Song Wang. Style-guided shadow removal. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XIX*, pages 361–378. Springer, 2022. 6, 7

[37] Jifeng Wang, Xiang Li, and Jian Yang. Stacked conditional generative adversarial networks for jointly learning shadow detection and shadow removal. In *CVPR*, pages 1788–1797, 2018. 1, 2, 6, 7

[38] Yufei Wang, Renjie Wan, Wenhan Yang, Haoliang Li, Lap-Pui Chau, and Alex Kot. Low-light image enhancement with normalizing flow. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 2604–2612, 2022. 2

[39] Yinhuai Wang, Jiwen Yu, and Jian Zhang. Zero-shot image restoration using denoising diffusion null-space model. *arXiv preprint arXiv:2212.00490*, 2022. 3

[40] Yufei Wang, Yi Yu, Wenhan Yang, Lanqing Guo, Lap-Pui Chau, Alex C Kot, and Bihan Wen. Exposurediffusion: Learning to expose for low-light image enhancement. *arXiv preprint arXiv:2307.07710*, 2023. 2

[41] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 6

[42] Zhendong Wang, Xiaodong Cun, Jianmin Bao, Wengang Zhou, Jianzhuang Liu, and Houqiang Li. Uformer: A general u-shaped transformer for image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17683–17693, 2022. 6

[43] Jay Whang, Mauricio Delbracio, Hossein Talebi, Chitwan Saharia, Alexandros G Dimakis, and Peyman Milanfar. Deblurring via stochastic refinement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16293–16303, 2022. 2

[44] Qingxiong Yang, Kar-Han Tan, and Narendra Ahuja. Shadow removal using bilateral filtering. 21(10):4361–4368, 2012. 6, 7

[45] Ling Zhang, Qing Zhang, and Chunxia Xiao. Shadow remover: Image shadow removal based on illumination recovering optimization. 24(11):4623–4636, 2015. 1, 2

[46] Wuming Zhang, Xi Zhao, Jean-Marie Morvan, and Liming Chen. Improving shadow suppression for illumination robust face recognition. 41(3):611–624, 2018. 1

[47] Lei Zhu, Zijun Deng, Xiaowei Hu, Chi-Wing Fu, Xuemiao Xu, Jing Qin, and Pheng-Ann Heng. Bidirectional feature pyramid network with recurrent attention residual modules for shadow detection. In *ECCV*, 2018. 6

[48] Lei Zhu, Zijun Deng, Xiaowei Hu, Chi-Wing Fu, Xuemiao Xu, Jing Qin, and Pheng-Ann Heng. Bidirectional feature pyramid network with recurrent attention residual modules for shadow detection. In *ECCV*, pages 121–136, 2018. 9

[49] Yurui Zhu, Jie Huang, Xueyang Fu, Feng Zhao, Qibin Sun, and Zheng-Jun Zha. Bijective mapping network for shadow removal. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5627–5636, 2022. 6, 7