ShadowDiffusion: Diffusion-based Shadow Removal using Classifier-driven Attention and Structure Preservation

Yeying Jin\(^1\), Wenhan Yang\(^2\), Wei Ye\(^3\), Yuan Yuan\(^3\) and Robby T. Tan\(^{1,4}\)

\(^1\)National University of Singapore, \(^2\)Peng Cheng Laboratory, \(^3\)Huawei International Pte Ltd, \(^4\)Yale-NUS College

jinyeying@u.nus.edu, yangwh@pcl.ac.cn, yewei10@huawei.com, yuanyuan10@huawei.com, robby.tan@nus.edu.sg

Abstract

Shadow removal from a single image is challenging, particularly with the presence of soft and self-shadows. Unlike hard shadows, soft shadows do not show any clear boundaries, while self-shadows are shadows that cast on the object itself. Most existing methods require the detection/annotation of binary shadow masks, without taking into account the ambiguous boundaries of soft and self-shadows. Most deep learning shadow removal methods are GAN-based and require statistical similarity between shadow and shadow-free domains. In contrast to these methods, in this paper, we present ShadowDiffusion, the first diffusion-based shadow removal method. ShadowDiffusion focuses on single-image shadow removal, even in the presence of soft and self-shadows. To guide the diffusion process to recover semantically meaningful structures during the reverse diffusion, we introduce a structure preservation loss, where we extract features from the pre-trained Vision Transformer (DINO-ViT). Moreover, to focus on the recovery of shadow regions, we inject classifier-driven attention into the architecture of the diffusion model. To maintain the consistent colors of the regions where the shadows have been removed, we introduce a chromaticity consistency loss. Our ShadowDiffusion outperforms state-of-the-art methods on the SRD, AISTD, LRSS, USR and UIUC datasets, removing hard, soft, and self-shadows robustly. Our method outperforms the SOTA method by 20\% of the RMSE of the whole image on the SRD dataset.

1. Introduction

Shadows can be categorized into hard shadows, soft shadows, and self-shadows \([21, 36]\). Hard and soft shadows, known as cast shadows, are formed when an object blocks beams of light and generates shadows on a surface or other objects nearby. Unlike hard shadows, soft shadows exhibit a gradual transition from shadow to non-shadow regions, without showing any sharp shadow boundaries. Self shadows are formed when part of an object blocks beams of light, casting shadows on the object itself. Generally, removing all these shadows is intractable due to ambiguities between shadows and non-shadows. However, compared with hard shadows, in most cases, soft and self shadows are more ambiguous and thus more difficult to identify and remove.

To remove shadows, some methods use an off-the-shelf shadow detection method \([50]\) or interactive user input \([13]\) to obtain binary shadow masks. However, these binary shadow masks are problematic to obtain, particularly for soft and self shadows. Moreover, due to the limited labeled shadow data or paired shadow and non-shadow images, the supervised shadow removal methods lack the generalisation beyond the relatively small training datasets.

Existing unsupervised methods (e.g., \([20, 23, 30]\)) are GAN-based \([49]\), which requires unpaired shadow and shadow-free images for training. They largely rely on
we introduce
ularly for self shadows and soft shadows. Moreover, these weakly-supervised methods require shadow-free regions cannot represent real-world shadow images. Therefore, once the two domains are statistically different, these methods produce hallucination/fake contents [50] and also suffer from unstable training.

The weakly-supervised methods (e.g., [28, 31]) are GAN-based and require binary shadow masks to distinguish shadow and shadow-free patches. These patches from the same images can reduce the domain gap of shadow and shadow-free domains. However, cropping patches is time-consuming [31] and synthesizing shadows on the shadow-free regions cannot represent real-world shadow images. Moreover, these weakly-supervised methods require shadows to be homogeneous, which is often not the case, particularly for soft shadows and soft shadows.

Unlike existing shadow removal methods, in this paper, we introduce ShadowDiffusion, a diffusion-based shadow removal method, to address hard, soft and self shadows in a single integrated framework. First, we constrain the shadow removal process by incorporating a classifier into denoising U-Net architecture in our diffusion model [18, 40]. By employing the binary classification of the shadow class and non-shadow class, we can inject shadow attention during the reverse diffusion iterations. Since our shadow attention is soft attention learned from the data, it can focus even on soft or self shadow regions.

Second, the structure and semantic information in shadow regions is often difficult to retain during the reverse diffusion, since the latent variables of DDPM (Denoising Diffusion Probabilistic Models [18, 39]) lack high-level semantic meaning [33]. To address this problem, namely, to guide the diffusion process to generate semantically meaningful structures, our basic idea is to inject the pre-trained feature extractor, DINO-ViT [2] into our ShadowDiffusion framework. Specifically, as the keys’ self-similarity [38] contains structure information [41], we use the structure preservation loss between intermediate keys of the clean and denoised images during the sampling. As shown in Fig. 1, our method can remove shadows and retain various structured regions (e.g., duck, paper, cleanser).

Third, our ShadowDiffusion is able to map shadow images to non-shadow images by learning from the training data. However, it has no explicit constraint on enforcing color consistency in the output images. To address this, we propose to utilize a normalized color (chromaticity) consistency loss. The key idea of this loss is to keep image chromaticity consistent before and after the shadow removal and independent from the light colors.

In summary, we make the following contributions:

1. We propose ShadowDiffusion, the first diffusion-based shadow removal network, that performs shadow removal robustly on hard, soft and self shadows from a single image in an integrated framework.

2. We propose a structure preservation loss based on pre-trained DiNO-ViT. This loss can preserve semantically meaningful structures in the reverse diffusion process of our shadow removal.

3. We introduce classification-driven attention in our diffusion model and chromaticity consistency loss. The classification-drive attention allows our diffusion model to focus on recovering shadow regions, including self and soft shadows. Our chromaticity-consistency loss enables our diffusion model to have consistent colors in our shadow removal outputs.

Comprehensive experiments on SRD, AISTD, LRSS, UIUC and USR datasets demonstrate that our ShadowDiffusion outperforms the state-of-the-art methods. It outperforms [51] by 14% error reduction (from 6.61 to 5.70) in terms of RMSE on shadow areas, 25% (from 3.61 to 2.72) on non-shadow areas and 20% (from 4.46 to 3.59) on the overall areas.

2. Related work

Different image priors have been explored for single image shadow removal, e.g., modeling of illumination and color [9], image regions [16,42], image gradients [13], user inputs [12]. However, they lack high-level semantics and may produce unsatisfactory results. In recent years, CNN-based shadow removal methods [3,10] show promising performance. DeshadowNet [34] removes shadows in an end-to-end manner. DSC [19] obtains global and context information from the direction-aware spatial attention module. SP+M-Net [27] and SP+M+I-Net [29] remove shadow using shadow image decomposition. CANet [3] is a two-stage context-aware network. BMNet [51] removes shadows from the perspective of invertible neural networks.

GAN-based shadow removal methods have also been studied. ST-CGAN [44] detects and removes shadows jointly using conditional GANs. ARGAN [6] has LSTM attention to detect shadows. DHAN [4] has MSGAN to generate shadow matting. RIS-GAN [46] explores the relationship between the residual images and the inverse illumination maps. SG-ShadowNet [43] treats shadow removal as intra-image style transfer. These supervised methods are Conditional GAN [22] or StyleGAN [25] based. Unsupervised shadow removal methods (e.g. Mask-ShadowGAN [20], LG-ShadowNet [30], DC-ShadowNet [23]) are CycleGAN-based [49], learn from unpaired data and exploit mask-guided cycle consistency constraints. Weakly supervised shadow removal methods wSP+M-Net [28] learn from shadow images, requiring shadow masks to distinguish shadow and non-shadow patches. G2R-ShadowNet [30] employs the shadow generators to produce pseudo shadow pairs for joint training. However, none of these methods can handle hard, soft
and self shadows in one unified framework. Recently, Denoising Diffusion Probabilistic Models [18, 39] (DDPM) and Denoising Diffusion Implicit Models (DDIM) [40] show promising generative capability. Unlike CNN-based and GAN-based methods, our ShadowDiffusion uses DDPM as the backbone, to address hard, soft and self shadows in a single integrated framework.

3. Proposed Method

Fig. 2 shows our pipeline, consisting of the forward diffusion and reverse denoising (generative) processes. The main goal of ShadowDiffusion is to remove hard, soft and self shadows using a diffusion-based network. To guide the diffusion model to focus on shadow regions, we propose classifier-driven attention, and inject the classifier into the U-Net noise estimator network. During the reverse diffusion, it is challenging to preserve the structures of the image. Therefore, we guide the reverse diffusion process using structure preservation loss.

3.1. Diffusion Models

We use Denoising Diffusion Probabilistic Models [18] as our generative model. DDPM is trained on a non-shadow image $x_0 \sim q(x_0)$ via a forward diffusion process $q(x_t \mid x_{t-1})$ that sequentially adds the Gaussian noise at every time steps $t$, i.e., $q(x_t \mid x_{t-1}) = \mathcal{N}(x_t; \sqrt{1 - \beta_t} x_{t-1}, \beta_t I)$, where $\beta_t = \frac{t}{T}$ is a variance schedule. The forward diffusion process follows a Markov chain of length $T$ expressed as:

$$q(x_{1:T} \mid x_0) = \prod_{t=1}^{T} q(x_t \mid x_{t-1}).$$

DDPM learns to reverse the process in Eq.(1) with the Gaussian transitions, i.e., $p_{\theta}(x_{t-1} \mid x_t) = \mathcal{N}(x_{t-1}; \mu_{\theta}(x_t, t), \Sigma_{\theta}(x_t, t))$, at each time step. The reverse denoising (generative) process is parameterized by a trainable network (e.g., U-Net), which estimates the mean $\mu_{\theta}(x_t, t)$ and variance $\Sigma_{\theta}(x_t, t)$ with parameter $\theta$. This reverse process starts from a standard normal distribution $p(x_T) = \mathcal{N}(x_T; 0, I)$ and follows:

$$p_{\theta}(x_0:T) = p(x_T) \prod_{t=1}^{T} p_{\theta}(x_{t-1} \mid x_t).$$

The DDPM model is optimized by considering the variational lower bound [5]. The noise prediction network $\epsilon_{\theta}(x_t, t)$ can be optimized by this objective function:

$$L_{\theta}(DM) = \mathbb{E}_{x_0, \epsilon, t} \left[ \| \epsilon - \epsilon_{\theta}(x_t, t) \|^2 \right].$$

After the optimization, we can sample from the learned parameterized Gaussian transitions $p_{\theta}(x_{t-1} \mid x_t)$ by:

$$x_{t-1} = \mu_{\theta}(x_t, t) + \Sigma_{\theta}(x_t, t) \epsilon, \quad \epsilon \sim \mathcal{N}(0, I).$$

Denoising Diffusion Implicit Models (DDIM) [40] shares the same objective in Eq. (3) as that of DDPM, yet has a deterministic and consistent reverse process. Our ShadowDiffusion use conditional DDIM to output deterministic and consistent samples.
Conditional Diffusion Models For the conditional diffusion models [35, 37], we conditionally reverse process \( p_\theta(x_{t-1} \mid x_t) \), without changing the diffusion process \( q(x_{1:T} \mid x_0) \) for \( x \) in Eq.(1). We replace the noise prediction network \( \epsilon_\theta(x_t, t) \) in Eq.(3) with \( \epsilon_\theta(\tilde{x}, x_t, t) \), where \( \tilde{x} \) denotes the spatial conditional image.

Here, we use the conditional DDIM [40] and with input \((\tilde{x}, x_t)\) to produce the output shadow-free image. The SRD [34] and AISTD [27] shadow datasets provide shadow images \( \tilde{x} \) and shadow-free images \( x_0 \), which are used in ShadowDiffusion training. We concatenate \( \tilde{x} \) and \( x \) channel-wisely, and input them to the deterministic reverse process. Therefore, we treat shadow removal as a reverse process with the loss function \( L_\theta(CDM) \) and parameterize \( f_\theta \) as a noise prediction network \( \epsilon_\theta(\tilde{x}, x_t, t) \):

\[
L_\theta(CDM) = E_{x_0, \epsilon, t} \left[ \| \epsilon - \epsilon_\theta(\tilde{x}, x_t, t) \|^2 \right], \tag{5}
\]

\[
f_\theta(\tilde{x}, x_t, t) = \frac{1}{\sqrt{\alpha_t}} \left( x_t - \sqrt{1 - \alpha_t} \cdot \epsilon_\theta(\tilde{x}, x_t, t) \right). \tag{6}
\]

The shadow removal starts from the noise map, \( x_T \sim \mathcal{N}(0, I) \), with the condition shadow input \( \tilde{x} \), and applies the diffusion towards the target non-shadow image \( x_0 \). The sampling from the reverse process \( x_{t-1} \sim p_\theta(x_{t-1} \mid x_t, \tilde{x}) \) employs:

\[
x_{t-1} = f_\theta(\tilde{x}, x_t, t) \sqrt{\alpha_{t-1}} + \epsilon_\theta(\tilde{x}, x_t, t) \sqrt{1 - \alpha_{t-1}}. \tag{6}
\]

3.2. Classifier-driven Attention

ShadowDiffusion employs time-conditional U-Net as the backbone, which has an encoder \( E \) and a decoder \( D \). To guide latent code \( z \) to focus on shadow regions during reverse sampling, we inject a classifier in between the encoder and decoder. The conditional latent diffusion itself has a cross-attention mechanism with query, key, value attention:

\[
\text{Attention}(Q, K, V) = \text{softmax} \left( \frac{QK^T}{\sqrt{d}} \right) \cdot V,
\]

where:

\[
Q = W_Q^{(i)} \cdot \varphi_i(z_t), \quad K = W_K^{(i)} \cdot \tau_\theta(y), \quad V = W_V^{(i)} \cdot \tau_\theta(y).
\]

In the equations, \( \varphi_i(z_t) \in \mathbb{R}^{N \times d_t^i} \) denotes a (flattened) intermediate representation of U-Net. The variables \( W_Q^{(i)} \in \mathbb{R}^{d \times d_t^i}, W_K^{(i)} \in \mathbb{R}^{d \times d_r}, W_V^{(i)} \in \mathbb{R}^{d \times d_r} \) are learnable projection matrices. We add a classifier into the encoder and decoder architecture, therefore injecting shadow attention during the reverse diffusion iterations. This attention mechanism is based on the binary classification of the shadow class and non-shadow class, by using CAM (class activation map) [47]. As shown in Fig. 3, we can obtain the activated attention to the shadow regions, particularly for soft and self shadow regions.

3.3. Structure Preservation Loss

Recent shadow removal methods do not consider the semantic information, which is strongly associated with the object structures, leading to unsatisfactory shadow removal results [34]. Moreover, during the reverse process of the diffusion model, it is challenging to maintain the original image structures. However, preserving structures is critical for shadow removal. For example, in Fig. 4, we want to maintain the overall structures of the dog, fire hydrant in the non-shadow regions, and also to recover the structures of bag, duck, and bear in the shadow regions. To preserve the structures, our key idea is to exploit a self-similarity prior [38]. For the soft shadow example in Fig. 4, the left and right parts of the bag contain different lighting conditions and appearances, but they do share similar structures (i.e., self-similarity).

A pre-trained DINo-ViT can provide more meaningful semantic features, compared to the CNN backbone (ResNet50), other transformer-based backbones (DeiT, ViT-Random), and a combination of them (ResNet50-DINO) shown in [24]. Motivated by this, we use a pre-trained DINo-ViT [2] as our feature extractor, enabling us to capture deep features [1], and providing useful feature supervi-
Figure 5. Our chromaticity consistency loss $L_{\text{chinf}}$ (grey). We calculate the corresponding projection direction $\alpha$ of the surface line using [7]. Using $\alpha$, we can obtain the light-invariant map and chromaticity map, which provide initial shadow-free guidance to ShadowDiffusion.

**DINO-ViT keys** We extract keys (deep features) from the pre-trained DINO-ViT. In the multi-head self-attention layer, the keys contain the structure information [41]. In Fig. 4, we show the Principal Component Analysis (PCA) visualization of the keys’ self-similarity and demonstrate the top three components as RGB at the deepest layer of DINO-ViT. As one can observe, the feature representations capture the object structures, helping our ShadowDiffusion preserves and recovers semantically meaningful structures.

We propose a structure preservation loss between intermediate keys of the non-shadow image and the estimated denoised output image during the reverse diffusion sampling:

$$L_{\text{stru}}(x, x_0) = \|S^l(x) - S^l(x_0)\|_F,$$

where $S$ is the self-similarity descriptor, $S^l(x) \in \mathbb{R}^{(n \times n)}$, with $n \times n$ dimension, where $n$ is the number of patches. In DINO-ViT, the input image is split into $n$ non-overlapping patches. $l = 11$ is the deepest ViT layer. $\| \cdot \|_F$ is the Frobenius norm. The self-similarity descriptor is defined as:

$$S^l(x)_{ij} = \cos(k^l_i(x), k^l_j(x)) = 1 - \frac{k^l_i(x) \cdot k^l_j(x)}{\|k^l_i(x)\| \cdot \|k^l_j(x)\|},$$

where $\cos(\cdot)$ is the cosine similarity between spatial keys $k_i$ and $k_j$.

### 3.4. Chromaticity Consistency Loss

In outdoor scenes, non-shadow regions are mostly lit by both the sunlight and skylight, while shadow regions are mostly lit by the skylight [8, 26]. Depending on the skylight when an image is captured, the light colors between the shadow and non-shadow regions might be different, even when the surface colors in both regions are the same. Ideally, the colors of the shadow regions after shadow removal should be consistent with those of the non-shadow regions of the same object. In other words, once we remove the shadows, we need to ensure that the recovered colors are the same as those under the sunlight and skylight. In indoor, this problem normally is not as significant as in outdoor, unless the inter-reflected light has a color that is considerably different from that of the ambient light.

In our framework, ShadowDiffusion learns how to diffuse input shadow images to our output shadow-free images. Besides learning the shadow characteristics, our ShadowDiffusion also learns how to transform the colors of shadow regions to those of non-shadow regions based on the initial non-shadow images. However, there is no explicit constraint that enforces color consistency before and after shadow removal. For this reason, we propose the use of a chromaticity consistency loss. The key idea of this loss is to keep image chromaticity consistent and independent from the light colors.

Finlayson et al. [7] introduce the log-chromaticity space where a surface color under different light colors can form a single line (the projection direction is $\alpha$) and thus can be projected into a single log-chromaticity value. Note, $\{\chi_1, \chi_2\}$ is the log-chromaticity value $\{\log(B/G), \log(R/G)\}$, where $R, G, B$ are the color channels. Based on this, we can enforce that a pixel’s color should have the same projected log-chromaticity value before and after shadow removal, which can be expressed as a light-invariant map $\chi_0 = \chi_1 \cos \alpha + \chi_2 \sin \alpha$, and chromaticity map $\chi_\alpha = P_\alpha \cdot \chi$, where $P_\alpha$ is a projector matrix. Our chromaticity consistency loss is defined as:

$$L_{\text{ch}}(x, x_0) = \|\chi_\alpha(x) - \chi_\alpha(x_0)\|_1,$$

Note that, for the self shadow example in Fig. 5 or other shadow examples in Fig. 6, which do not have the paired clean images $x_0$, used in training, our chromaticity consistency loss can work during inference. The chromaticity consistency used in inference is defined as:

$$L_{\text{chinf}}(x, \tilde{x}) = \|\chi_\alpha(x) - \chi_\alpha(\tilde{x})\|_1.$$
4. Experiments

To ensure fair comparisons in our experiments, all the baselines, including ours are trained and tested on the same datasets. We trained our ShadowDiffusion on each dataset and tested on the corresponding dataset, e.g., for SRD [34], we used 2680 SRD and 408 SRD images for training and testing, respectively. Moreover, lines, including ours are trained and tested on the same datasets.

### Baselines

We include SOTA supervised methods SG-ShadowNet’22 [43], BMNet’22 [51], SP+M+I-Net [29]; the SOTA weakly supervised method G2R-ShadowNet [30]; and the SOTA unsupervised method DC-ShadowNet [23]. We also compared ShadowDiffusion to the traditional method Guo [17].

### 4.1. Shadow Removal on Hard Shadows

The SRD dataset consists of 2680 training and 408 testing pairs of shadow and shadow-free images without shadow masks. It contains hard shadows and soft shadows. For the baseline methods that require shadow masks, we additionally use the processing results of DHAN [4] for SRD shadow masks.

Fig. 7 (bottom four rows) and Table 1 show the hard shadow removal results on the SRD dataset. Fig. 8 (bottom three rows) show the soft shadow removal results on the SRD dataset, which show our ShadowDiffusion outperforms the SOTA methods. The best and the second-best values for each metric are highlighted in bold and underlined, respectively. Our ShadowDiffusion presents a competitive shadow removal performance by outperforming [51] by 14% error reduction (from 6.61 to 5.70) in terms of RMSE on shadow areas, 25% (from 3.61 to 2.72) on non-shadow areas and 20% (from 4.46 to 3.59) on the overall areas. Moreover,

#### Table 1. Quantitative results on the SRD dataset. S, NS and ALL represent shadow, non-shadow and entire regions, respectively. M shows that ground truth shadow masks are also used in training.

| Methods   | Train | S     | NS    | ALL  | S     | NS    | ALL  | S     | NS    | ALL  |
|-----------|-------|-------|-------|------|-------|-------|------|-------|-------|------|
| DC [23]   | UP    | 3.59  | 3.05  | 2.77 | 0.975 | 0.981 | 0.955|
| DHAN [4]  | P+M   | 8.41  | 6.80  | 7.79 | 0.978 | 0.979 | 0.957|
| Auto [10] | P+M   | 8.56  | 6.75  | 7.56 | 0.978 | 0.979 | 0.956|
| EM’22 [52]| P+M   | 10.00 | 7.20  | 6.04 | 0.937 | 0.879 | 0.779|
| BM’22 [51]| P+M   | 6.61  | 4.46  | 3.30 | 0.981 | 0.982 | 0.956|
| SG’22 [43]| P+M   | 5.73  | 3.08  | 4.33 | 0.981 | 0.987 | 0.959|
| Ours      | P     | 5.79  | 3.78  | 3.27 | 0.984 | 0.988 | 0.968|

The reason is from the input shadow image \( x \) and the corresponding projection direction \( \alpha \), using the surface line constraint, the light-invariant map and the chromaticity map provide the initial shadow-free guidance (see Fig. 5) to the ShadowDiffusion network. This constraint helps enforce the consistency of the colors/chromaticities of the pixels, so that the colors are at least not transformed into arbitrary colors by our network.

#### Table 2. RMSE results (lower is better) on the AISTD dataset.

| Methods   | Train | All | S     | NS   |
|-----------|-------|-----|-------|------|
| Input Image | -     | 8.5 | 40.2  | 2.6  |
| MaskSGAN [20] | UP    | 5.3 | 12.5  | 4.0  |
| DC [23]    | UP    | 4.6 | 10.3  | 3.5  |
| LG [30]    | UP    | 4.4 | 9.9   | 3.4  |
| Yang et al. [45] | -     | 16.0 | 24.7  | 14.4 |
| Gong et al. [11] | -    | 13.3 | -     | -    |
| wSP+M-Net [28] | M   | 4.1 | 9.7   | 3.0  |
| G2R [30]   | M     | 3.9 | 8.8   | 2.9  |
| Guo et al. [14] | P+M  | 6.1 | 22.0  | 3.1  |
| ST-CGAN [44] | P     | 8.7 | 13.4  | 7.9  |
| DHAN [4]   | P+M   | 7.9 | 11.4  | 7.2  |
| Auto [10]  | P+M   | 4.2 | 6.5   | 3.8  |
| SP+M-Net [27] | P     | 3.9 | 7.9   | 3.1  |
| SP+M+I-Net [29] | P+M | 3.6 | 6.0   | 3.1  |
| BMNet’22 [51]| P+M  | 3.5 | 6.1   | 2.9  |
| SG’22 [43]| P+M   | 3.5 | 6.0   | 3.1  |
| Deshadow [34] | P   | 7.6 | 15.9  | 6.0  |
| SP+M-Net [27] | P   | 4.0 | 8.2   | 3.3  |
| SP+M+I-Net [29] | P  | 4.0 | 6.6   | 3.5  |
| Ours      | P     | 3.6 | 5.9   | 2.9  |

---

1. As mentioned in [32], the RMSE evaluation used by all methods computes MAE (mean absolute error).

2. The evaluation codes adopt from [10].

---

Color space between the shadow remove the output and its ground truth. Following [10], we calculated the peak signal-to-noise ratio (PSNR) and structural similarity (SSIM) in the RGB color space; the higher, the better.
ShadowDiffusion achieves the highest PSNR and SSIM values in shadow, non-shadow and overall areas.

The AISTD dataset contains 1870 image triplets (shadow images, shadow-free images, and shadow masks), with 1330 triplets for training and 540 triplets for testing. The shadows in AISTD dataset are mainly hard shadows. The AISTD dataset (A means adjusted) reduced the color discrepancy between shadow and shadow-free images of the original ISTD dataset [44], thus is more reliable to use in training. Fig. 7 (second row) and Table 2 show results on the AISTD dataset, which demonstrate the superiority of our ShadowDiffusion compared to baseline results in terms of RMSE in shadow regions and non-shadow regions.

The USR dataset [20] is an unpaired dataset, including hard, soft, and self shadows. We evaluate on USR test shadow images and show results in Fig. 7 (first row).

4.2. Shadow Removal on Soft and Self Shadows

The LRSS [13] is a soft shadow dataset with 134 shadow images\(^3\); we followed [13, 23], using the same 34 LRSS images with their corresponding shadow-free images for evaluation shown in Table 3. Our ShadowDiffusion achieves the highest PSNR and lowest RMSE.

Fig. 8 (top three rows) show results on the LRSS dataset. Besides the SOTA baselines, we compared to the traditional hard and soft shadow removal method [16]. Since the majority of the SOTA baselines need the shadow mask for evaluation, we use the shadow mask provided by LRSS dataset [13]. However, the provided binary shadow mask does not consider the ambiguous boundaries of soft and self shadows, thus leading to remaining shadows in the ambiguous boundaries.

UCF [48] and UIUC dataset [15], contain 245 and 108 images. UIUC provides 30 images where shadows are caused by objects in the scene, which are used for self shadow evaluation, the results are shown in Fig. 9. For the baseline methods require shadow masks, we additionally use the detection results of BDRAR [50].

4.3. Ablation Studies

Fig. 10 and Table 4 show the effectiveness of classifier-driven attention, structure preservation loss and chromaticity loss used in ShadowDiffusion. For the ablation studies of classifier-driven attention shown in Fig. 10 (first row), we remove the classifier in the U-Net architecture and can observe the remaining shadows. The activated shadow attention features are shown in Fig. 3. From Fig. 10 (second row), the structure preservation loss help in maintaining the structures of the mountain, owing to the structure fea-

---

\(^3\)LRSS Dataset are obtained from their project website: http://visual.cs.ucl.ac.uk/pubs/softshadows/
5. Conclusion

In this paper, we have proposed ShadowDiffusion, the first conditional diffusion-based shadow removal framework. We inject classifier-driven attention to our ShadowDiffusion framework, so that our ShadowDiffusion can focus on hard, soft and self shadow regions. To guide our diffusion process to preserve and recover structure information, we propose structure preservation loss. Moreover, our chromaticity consistency loss keeps image chromaticity consistent and helps recover the color of shadow regions. Unlike existing methods, our ShadowDiffusion performs shadow removal robustly on hard, soft and self shadows. Our experiments show our method outperforms a SOTA method [51] by 14% error reduction (from 6.61 to 5.70) in terms of RMSE on
shadow areas, 25% (from 3.61 to 2.72) on non-shadow areas and
20% (from 4.46 to 3.59) on the overall areas.

Acknowledgment

This research/project is supported by the National Research
Foundation, Singapore under its AI Singapore Programme (AISG
Award No: AISG2-PhD/2022-01-037[T]), and partially supported
by MOE2019-T2-1-130. Robby T. Tan’s work is supported by
MOE2019-T2-1-130.

References

[1] Shir Amir, Yossi Gandelsman, Shai Bagon, and Tali Dekel.
Deep vit features as dense visual descriptors. arXiv preprint
arXiv:2112.05814, 2021.

[2] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou,
Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging
properties in self-supervised vision transformers. In Proceed-
ing of the IEEE/CVF International Conference on Com-
puter Vision, pages 9650–9660, 2021.

[3] Zipei Chen, Chengjiang Long, Ling Zhang, and Chunxiao
Xiao. Canet: A context-aware network for shadow removal.
In Proceedings of the IEEE/CVF International Conference on
Computer Vision, pages 4743–4752, 2021.

[4] Xiaodong Cun, Chi-Man Pun, and Cheng Shi. Towards
ghost-free shadow removal via dual hierarchical aggregation
network and shadow matting gan. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 34,
pages 10680–10687, 2020.

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

[6] Bin Ding, Chengjiang Long, Ling Zhang, and Chunxiao Xiao.
Argan: Attentive recurrent generative adversarial network
for shadow detection and removal. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 10213–10222, 2019.

[7] Graham D Finlayson, Mark S Drew, and Cheng Lu. Entropy
minimization for shadow removal. International Journal of
Computer Vision, 85(1):35–57, 2009.

[8] Graham D Finlayson, Brian V Funt, and Kobus Barnard.
Color constancy under varying illumination. In Proceed-
ings of IEEE International Conference on Computer Vision,
pages 720–725. IEEE, 1995.

[9] Graham D Finlayson, Steven D Hordley, Cheng Lu, and
Mark S Drew. On the removal of shadows from images.
IEEE transactions on pattern analysis and machine intel-
ligence, 28(1):59–68, 2005.

[10] 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 Proceed-
ings of the IEEE/CVF Conference on Computer Vision and
Pattern Recognition, pages 10571–10580, 2021.

[11] Han Gong and Darren Cosker. Interactive shadow removal
and ground truth for variable scene categories. In BMVC,
pages 1–11, 2014.

[12] Han Gong and Darren Cosker. Interactive removal and
ground truth for difficult shadow scenes. Journal of the Opt-
tical Society of America A, 33(9):1798–1811, 2016.

[13] Maciej Gryka, Michael Terry, and Gabriel J Brostow. Learn-
ing to remove soft shadows. ACM Transactions on Graphics
(TOG), 34(5):1–15, 2015.

[14] Ruiqi Guo, Qiyeun Dai, and Derek Hoiem. Single-image
shadow detection and removal using paired regions. In Pro-
ceedings of the IEEE/CVF Conference on Computer Vision
and Pattern Recognition, pages 2033–2040, 2011.

[15] Ruiqi Guo, Qiyeun Dai, and Derek Hoiem. Single-image
shadow detection and removal using paired regions. In Pro-
ceedings of the IEEE Conference on Computer Vision and
Pattern Recognition, pages 2033–2040, 2011.

[16] Ruiqi Guo, Qiyeun Dai, and Derek Hoiem. Paired regions for
shadow detection and removal. IEEE transactions on pattern
analysis and machine intelligence, 35(12):2956–2967, 2012.

[17] Ruiqi Guo, Qiyeun Dai, and Derek Hoiem. Paired regions for
shadow detection and removal. IEEE transactions on pattern
analysis and machine intelligence, 35(12):2956–2967, 2012.

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

[19] Xiaowei Hu, Chi-Wing Fu, Lei Zhu, Jing Qin, and Pheng-
Ann Heng. Direction-aware spatial context features for
shadow detection and removal. IEEE Transactions on Pattern
Analysis and Machine Intelligence, 42(11):2795–2808,
2019.

[20] Xiaowei Hu, Yitong Jiang, Chi-Wing Fu, and Pheng-Ann
Heng. Mask-shadowgan: Learning to remove shadows from
unpaired data. In Proceedings of the IEEE International
Conference on Computer Vision, pages 2472–2481, 2019.

[21] Jia-Bin Huang and Chu-Song Chen. Moving cast shadow
detection using physics-based features. In 2009 IEEE Con-
ference on Computer Vision and Pattern Recognition, pages
2310–2317. IEEE, 2009.

[22] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A
Efros. Image-to-image translation with conditional adversar-
ial networks. In Proceedings of the IEEE/CVF Conference on
Computer Vision and Pattern Recognition, pages 1125–
1134, 2017.

[23] Yeying Jin, Aashish Sharma, and Robby T Tan. Dc-
shadownet: Single-image hard and soft shadow removal using
unsupervised domain-classifier guided network. In Proceed-
ings of the IEEE/CVF International Conference on Com-
puter Vision, pages 5027–5036, 2021.

[24] Tejan Karmali, Rishubh Parihar, Susmit Agrawal, Harsh
Rangwani, Varun Jampani, Maneesh Singh, and R Venkatesh
Babu. Hierarchical semantic regularization of latent spaces in
stylegans. In European Conference on Computer Vision,
pages 443–459. Springer, 2022.

[25] Tero Karras, Samuli Laine, and Timo Aila. A style-based
generator architecture for generative adversarial networks.
In Proceedings of the IEEE/CVF conference on computer vi-
sion and pattern recognition, pages 4401–4410, 2019.
[26] Rei Kawakami, Katsushi Ikeuchi, and Robby T Tan. Consistent surface color for texturing large objects in outdoor scenes. In *Proceedings of IEEE International Conference on Computer Vision*, volume 2, pages 1200–1207. IEEE, 2005.

[27] Hieu Le and Dimitris Samaras. Shadow removal via shadow image decomposition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8578–8587, 2019.

[28] Hieu Le and Dimitris Samaras. From shadow segmentation to shadow removal. In *European Conference on Computer Vision*, pages 264–281. Springer, 2020.

[29] Hieu Le and Dimitris Samaras. Physics-based shadow image decomposition for shadow removal. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (01):1–1, 2021.

[30] 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.

[31] Zhihao Liu, Hui Yin, Xinyi Wu, Zhenyao Wu, Yang Mi, and Song Wang. From shadow generation to shadow removal. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4927–4936, 2021.

[32] Official Github page of 'Shadow Removal via Shadow Image Decomposition' ICCV 2019. https://github.com/cvlab-stonybrook/SID.

[33] Konpat Preechakul, Nattanat Chatthee, Sutitisak Wizadwongsa, and Supasorn Suwajanakorn. Diffusion autoencoders: Toward a meaningful and decodable representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10619–10629, 2022.

[34] Liangqiong Qu, Jiandong Tian, Shengfeng He, Yandong Tang, and Rynson WH Lau. Deshadownet: A multi-context embedding deep network for shadow removal. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4067–4075, 2017.

[35] Chitwan Saharia, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. In *ACM SIGGRAPH 2022 Conference Proceedings*, pages 1–10, 2022.

[36] Elena Salvador, Andrea Cavallaro, and Touradj Ebrahimi. Cast shadow segmentation using invariant color features. *Computer vision and image understanding*, 95(2):238–259, 2004.

[37] Hiroshi Sasaki, Chris G Willcocks, and Toby P Breckon. Unit-ddpm: Unpaired image translation with denoising diffusion probabilistic models. *arXiv preprint arXiv:2104.05358*, 2021.

[38] Eli Shechtman and Michal Irani. Matching local self-similarities across images and videos. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2007.

[39] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pages 2256–2265, 2015.

[40] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021.

[41] Narek Tumanyan, Omer Bar-Tal, Shai Bagon, and Tali Dekel. Splicing vit features for semantic appearance transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10748–10757, 2022.

[42] Tomas F Yago Vicente, Minh Hoai, and Dimitris Samaras. Leave-one-out kernel optimization for shadow detection and removal. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(3):682–695, 2017.

[43] Jinfan Wan, Hui Yin, Zhenyao Wu, Xinyi Wu, Yanting Liu, and Song Wang. Style-guided shadow removal. In *European Conference on Computer Vision*, pages 361–378. Springer, 2022.

[44] Jifeng Wang, Xiang Li, and Jian Yang. Stacked conditional generative adversarial networks for jointly learning shadow detection and shadow removal. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1788–1797, 2018.

[45] Qingxiong Yang, Kar-Han Tan, and Narendra Ahuja. Shadow removal using bilateral filtering. *IEEE Transactions on Image Processing*, 21(10):4361–4368, 2012.

[46] Ling Zhang, Chengjiang Long, Xiaolong Zhang, and Chunxia Xiao. Ris-gan: Explore residual and illumination with generative adversarial networks for shadow removal. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12829–12836, 2020.

[47] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[48] Jiejie Zhu, Kegan GG Samuel, Syed Z Masood, and Marshall F Tappen. Learning to recognize shadows in monochromatic natural images. In *2010 IEEE Computer Society conference on computer vision and pattern recognition*, pages 223–230. IEEE, 2010.

[49] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2223–2232, 2017.

[50] 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 *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 121–136, 2018.

[51] Yurui Zhu, Jie Huang, Xiaoyang 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.
[52] Yurui Zhu, Zeyu Xiao, Yanchi Fang, Xueyang Fu, Zhiwei Xiong, and Zheng-Jun Zha. Efficient model-driven network for shadow removal. 2022. 6