Semantic-guided Coarse-to-Fine Diffusion Model for Self-supervised Image Shadow Removal

Ziqi Zeng
Nanjing Normal University
Nanjing, Jiangsu,China

Chen Zhao
Nanjing University
Nanjing, Jiangsu,China

Weiling Cai
Nanjing Normal University
Nanjing, Jiangsu,China
caiwl@njnu.edu.cn

Yuqing Guo
Nanjing Normal University
Nanjing, Jiangsu,China

Abstract

Existing unsupervised methods have addressed the challenges of inconsistent paired data and tedious acquisition of ground-truth labels in shadow removal tasks. However, GAN-based training often faces issues such as mode collapse and unstable optimization. Furthermore, due to the complex mapping between shadow and shadow-free domains, merely relying on adversarial learning is not enough to capture the underlying relationship between two domains, resulting in low quality of the generated images. To address these problems, we propose a semantic-guided coarse-to-fine diffusion model for self-supervised shadow removal, which consists of two stages. In the first stage, a semantic-guided generative adversarial network (SG-GAN) is proposed to carry out a coarse result and construct paired synthetic data through a cycle-consistent structure. Then the coarse result is refined with a diffusion-based restoration module (DBRM) to enhance the texture details and edge artifact at second stage. Meanwhile, we propose a multi-modal semantic prompter (MSP) that aids in extracting accurate semantic information from real images and text, guiding the shadow removal network to restore images better in SG-GAN. We conduct experiments on multiple public datasets and the experimental results demonstrate the effectiveness of our method.

Keywords: Shadow removal, Semantic guidance, Diffusion model.

1. Introduction

Shadows, a natural consequence of obstructed light sources, play a crucial role in shaping the visual landscape. Although shadows convey valuable signals regarding object shapes and light direction, their presence often complicates the semantic understanding of images in computer vision tasks such as image segmentation [25] and object detection [7]. Shadow regions may be misclassified as objects or as parts of objects, which can significantly impair the accuracy and performance of these tasks. Consequently, shadow detection and removal are essential for enhancing the efficacy of computer-based visual tasks.

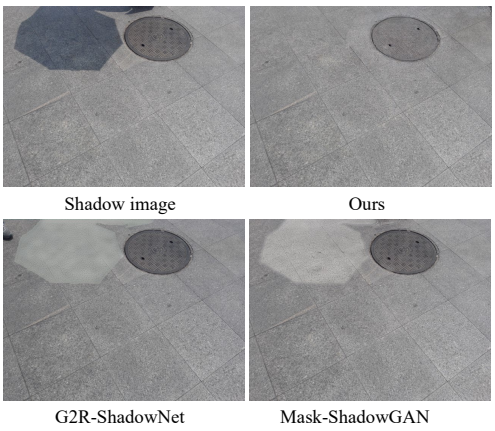


Figure 1. The shadow removal results of our method and other two GAN-based methods: G2R-ShadowNet[28] and Mask-ShadowGAN[15]. GAN-based methods have obvious shadow boundaries and artifacts.

Early shadow removal methods [2, 8, 12, 44]focus on exploring shadow removal in images based on different physical properties of shadows. Due to the insufficient accuracy and limitations of the underlying physical model, traditional physical model-based shadow removal algorithms are unable to effectively address shadows in complex real-world scenes [26].

Learning-based methods [5, 47, 52] typically train networks using paired shadow images and corresponding

shadow-free images in a fully supervised manner. However, inconsistencies exist in large-scale paired shadow removal datasets due to the uncontrollable nature of outdoor illumination [22]. Some methods have been proposed that do not require paired data for training and generate supervisory signals through shadow generation using a cycle-consistent architecture. Nevertheless, the gap between synthetic and real images limits the full application of these methods to real-world shadow removal tasks. Although existing unsupervised methods have achieved notable results in shadow removal, GAN-based methods are susceptible to unstable optimization and mode collapse during training [45, 49] Additionally, adversarial training alone is insufficient to fully learn the complex mapping between shadow and shadow-free domains. As shown in Fig. 1, the resulting shadow-free images often exhibit visible boundary artifacts and lack detailed texture restoration, leaving room for improvement in visual quality.

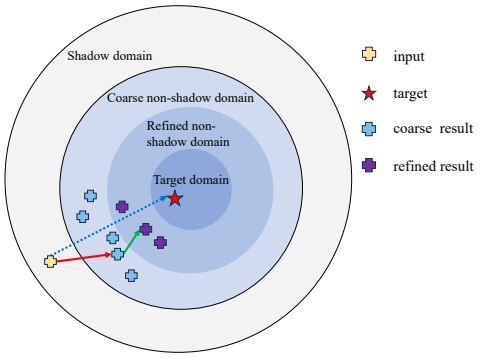


Figure 2. Dagram of different domains and domain transitions. The red arrow represents the previous adversarial generation process, and the green arrow represents the diffusion generation process.

In recent years, diffusion models, e.g., the Denoising Diffusion Probabilistic Model (DDPM) [13], have achieved significant breakthroughs in visual generative tasks as a new branch of generative models [24, 30, 31, 38, 39, 50, 51]. Owing to their superior generative capabilities, many studies have explored the application of diffusion models in image restoration tasks to enhance texture recovery. ShadowDiffusion [10] proposed a diffusion sampling strategy to explicitly integrate the shadow degradation prior into the inherent iterative process of the dynamic mask sensing diffusion model. Liu [26] reconstructed the local illumination of the shadow region using a diffusion model. Although diffusion models provide a more stable training process and are more effective in capturing the pixel distribution of images in contrast to GANs, these diffusion-based methods still require paired data for training and are encumbered by the limitations inherent to the supervised paradigm. Consequently, we aim to explore a self-supervised paradigm based on diffusion models.

As shown in Fig. 2, we first consider using a GAN-based method to carry out coarse shadow removal, and during this process, construct paired synthetic data through a cycle-consistent structure. Then we use a diffusion model to process the paired data and refine the coarse results to bring them closer to our target. Through this process, the diffusion model can be applied in self-supervised shadow removal. In this paper, we propose a novel semantic-guided coarse-to-fine diffusion model, which provides a self-supervised solution to shadow removal with unpaired data. Specifically, our framework comprises two major stages: a coarse processing stage and a refined restoration stage, which correspond to the semantic-guided generative adversarial network (SG-GAN) and the diffusion-based restoration module (DBRM), respectively. In the first stage, SG-GAN consists of two sub-branches. We train mask-guided generators to synthesize shadow images, assisting the shadow removal network in adversarial training. Then, we propose a multi-modal semantic prompter (MSP) that uses pre-trained visual-language models to extract features and semantic information from real images and text, enhancing shadow removal performance in both branches. In the second stage, we use paired data as the input, which is constructed by the cycle-consistent structure in the SG-GAN branch. We exploit DBRM to further refine the coarse results obtained from SG-GAN, which might contain edge artifacts and texture blurring, making them closer to the target images. This process overcomes the obstacle that the diffusion model relies on paired data for training, which otherwise makes its use in unsupervised methods challenging.

In summary, our main contributions of this work are as follows:

- We propose a semantic-guided coarse-to-fine diffusion model for self-supervised shadow removal to solve the difficulty of diffusion model processing unpaired data by constructing paired images with a cycle-consistent structure. Our methods can learn to remove shadow from unpaired data and solve the problem of obvious shadow boundary and texture detail missing in results.

- A general-purpose multi-modal semantic prompter is introduced to bridge the inherent gap between real-world shadow images and synthetic shadow images. Meanwhile, the effectiveness of this module is validated in other methods.

- We conduct extensive experiments on four public datasets. The experimental results show that our proposed method achieves competitive performance and is superior to previous unsupervised shadow removal methods.

2. Related work

2.1. Shadow removal

Traditional shadow removal methods rely on image gradients [8], illumination information [44], and image intensity regions [12] to remove shadows. These early shadow removal methods often model the image without shadows or transfer color and texture features from non-shadow regions to shadow regions to achieve shadow removal. However, due to the lack of accuracy in the underlying physical models, these methods usually cannot handle shadows in complex real-world scenes. With the emergence of deep learning methods, deep learning-based approaches have demonstrated greater advantages in dealing with more complex and varied scenes.

Qu [36] proposed an end-to-end deep neural network called DeshadowNet for shadow removal, which predicts shadow-free outputs based on three different directional inputs using multiple contextual architectures. Wang [43] designed ST-CGAN to detect and remove shadows and created the first large-scale shadow benchmark dataset, ISTD, consisting of 1870 pairs of shadowed and shadow-free images. Due to the difficulty and inconsistency in obtaining paired shadow images in practice, Hu [15] proposed the Mask-ShadowGAN method, based on the idea of CycleGAN, treating the shadow removal problem as an image-to-image style transfer task. LG-ShadowNet [27] proposed a shadow image enhancement method based on a simple physical lighting model and an image decomposition formula for shadow and pseudo-shadow removal. Liu [28] introduced G2R-ShadowNet for shadow removal using a training dataset constructed from shadow images and their corresponding shadow masks. Most of these methods adopt adversarial learning, which often results in noticeable shadow boundaries and a lack of texture details.

2.2. Diffusion model

Diffusion models [13, 41] are generative models that learn the distribution of real images through a Gaussian noise blurring process and a reverse denoising process. They have been successfully applied to various computer vision tasks, such as image super-resolution [40], inpainting [32], color harmonization [46], and image restoration [11, 48].

In recent years, diffusion models have also been used in shadow removal tasks. For example, [34] enhanced the diffusion process by conditioning on a learned latent feature space from shadow-free images while integrating noise features to avoid local optima during training. Liu [26] used diffusion models to reconstruct the local illumination of the shadow region based on the global illumination of the shadow image. However, these methods inevitably require paired data to provide supervisory information for network training.

Additionally, methods combining adversarial learning with diffusion models have emerged. For instance, [19] employed DDPM with adversarial learning for unsupervised vessel segmentation and achieved promising results. However, to the best of our knowledge, no study has yet combined these approaches in the domain of shadow removal.

3. Preliminary

In this section, we briefly review the key concepts underlying SDE-based diffusion models and outline the process of generating samples with reverse-time SDEs. Let p_0 denote the initial data distribution, and let $t \in [0, T]$ denote the continuous time variable. We consider a diffusion process $x(t)_{t=0}^{T}$ defined by an SDE of the form:

$$dx = f(x,t)dt + g(t)dw, \quad x(0) \sim p_0(x), \tag{1}$$

where f and g are the drift and dispersion functions, respectively, w is a standard Wiener process, and $x(0) \in \mathbb{R}^d$ is the initial condition. Typically, the terminal state $x(T)$ follows a Gaussian distribution with fixed mean and variance. The goal is to design such an SDE that gradually transforms the data distribution into Gaussian noise [6, 29].

We can reverse the process to sample data from noise by simulating the SDE backward in time [42]. [1] shows that a reverse-time representation of Eq.1 is given by:

$$dx = \left[f(x,t) - g(t)^2 \nabla_x \log p_t(x)\right] dt + g(t)d\hat{w}, \tag{2}$$

where $x(T) \sim p_T(x)$. Here, $\hat{w}$ is a reverse-time Wiener process, and $p_t(x)$ is the marginal probability density function of $x(t)$ at time t. Since the score function $\nabla_x \log p_t(x)$ is generally intractable, SDE-based diffusion models approximate it by training a time-dependent neural network $s_\theta(x,t)$ using a score matching objective [42].

4. Proposed method

4.1. Overall framework

To enhance the effectiveness of shadow removal networks and address edge artifacts and blurred textures in unsupervised methods, we propose a semantic-guided adversarial diffusion model for self-supervised shadow removal. Fig. 3 illustrates the overall network architecture of our method, which consists of two stages: the coarse processing stage and the refined restoration stage. These two stages are composed of the semantic-guided generative adversarial network (SG-GAN) and the diffusion-based restoration module (DBRM), respectively. In SG-GAN, we use a general-purpose multi-modal semantic prompter (MSP) to extract semantic information from a pre-trained CLIP model, which helps the network improve restoration. Structurally, SG-GAN is divided into two branches that

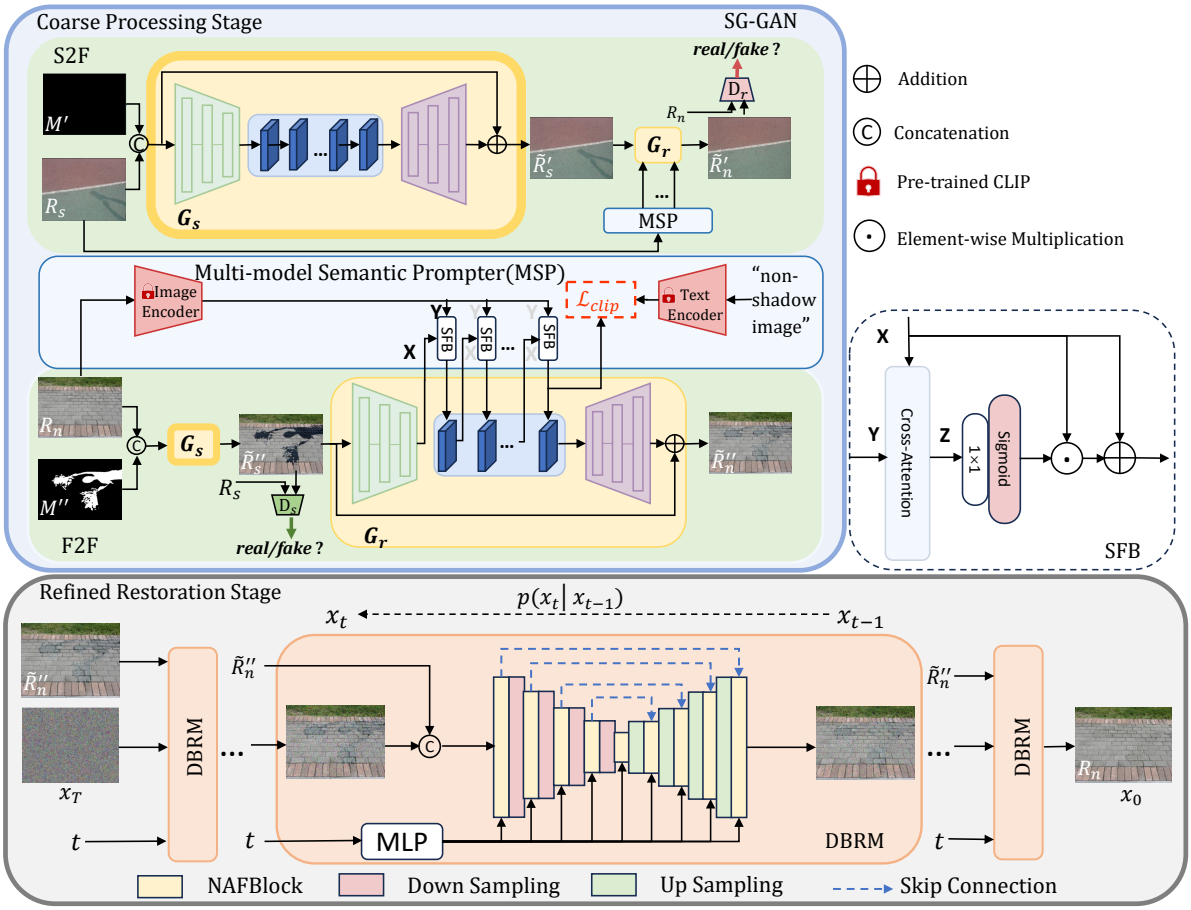


Figure 3. Overall pipeline of our method. At the coarse processing stage, SG-GAN, which consists of S2F, F2F, and MSP, predicts the coarse shadow removal results $\widetilde{R}_n''$. At the refined restoration stage, DBRM takes paired data R_n and $\widetilde{R}_n''$ from the previous stage as input, where the coarse result $\widetilde{R}_n''$ is refined.

utilize a set of unpaired shadow images, shadow-free images, and shadow masks as inputs for training. It includes shadow generation and removal generators G_s and G_r, as well as shadow and shadow-free image discriminators D_s and D_r. However, SG-GAN relies on discriminators and consistency constraints, which are insufficient to achieve optimal results. The outcomes after shadow removal still contain undesirable noise. Therefore, in the refined restoration stage, DBRM uses the coarse shadow removal results obtained from SG-GAN to construct paired data with clean inputs for network training. Using the powerful generative capabilities of the diffusion model, DBRM further refines the coarse results to remove artifacts and improve texture details.

4.2. Multi-modal semantic prompter

Existing unsupervised methods [15, 28] primarily rely on generated shadow images (synthetic data) to train shadow removal networks without guidance from real prior information. However, as shown in Fig. 4, even though the real shadow image and the synthesized shadow image look very similar, there is still a difference in their data distribution. The shadow removal network trained with synthetic images might see a reduction in effectiveness on real-world shadow images.

Useful prompts can help correct task networks for better performance. To minimize the impact of this gap, we designed a multi-modal semantic prompter (MSP). As shown in Fig. 3, MSP extracts features using the image encoder C_{image} and the text encoder C_{text} from a pre-trained CLIP model [37]. The prior information extracted by the image encoder is fused with the features extracted by G_r in residual blocks through semantic fusion blocks (SFB), while the text features extracted by the text encoder are used to define a contrastive loss (introduced in Sec.4.3) to further constrain the image recovered by G_r.

SFB aims to better perceive prior information, using this more reliably perceived content to assist G_r in restoration.

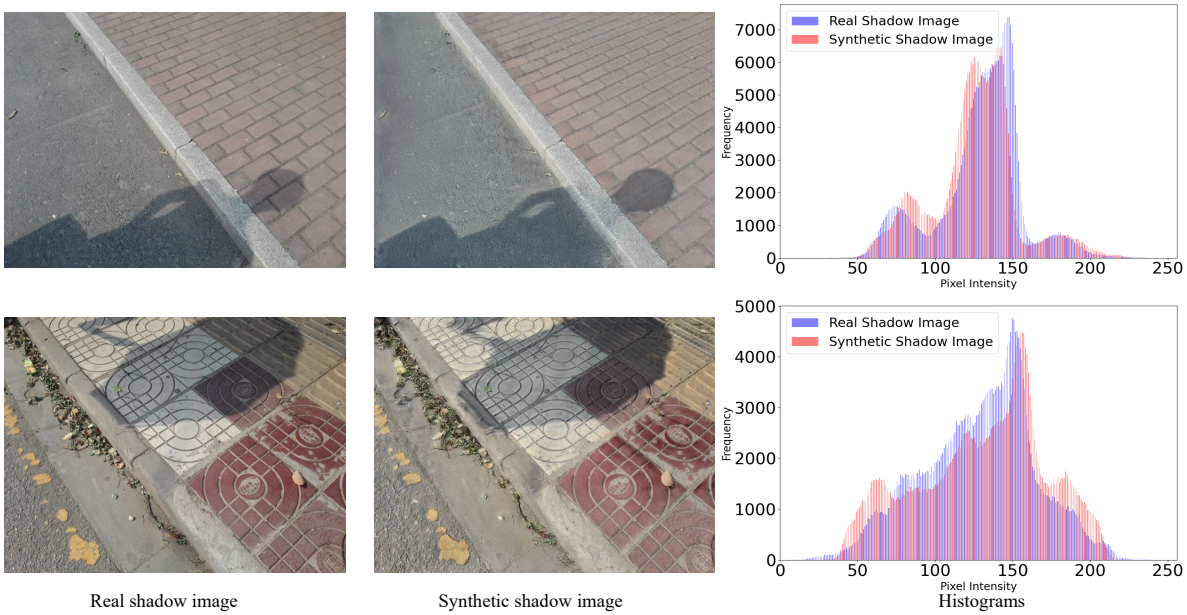


Figure 4. Real shadow images from ISTD and corresponding synthetic shadow images obtained from our generator G_s, the histograms show the inconsistencies in the intensity distribution between them.

It also controls the propagation of perceived prior information, enabling the network to adaptively learn more useful features in G_r for better restoration. Given a recovery tensor X_{k-1} from the $(k-1)^{th}$ residual block in G_r and a semantic tensor Y extracted by C_{image}, they are first fused through a cross-attention mechanism to obtain Z_{k-1}:

$$Z_{k-1} = A_{c-attn}(X_{k-1}, Y), \tag{3}$$

where A_{c-attn} refers to the cross-attention mechanism [16]. Then, using a 1×1 convolution and a gated control function, sigmoid, the input for the next residual block can be obtained:

$$\mathcal{S}(X_{k-1}, Y) = \sigma(W_p(Z_{k-1})) \odot X_{k-1} + X_{k-1}, \tag{4}$$

where $\sigma(\cdot)$ represents the sigmoid function and $W_p(\cdot)$ represents the 1×1 point-wise convolution.

4.3. Semantic-guided generative adversarial network

SG-GAN consists of two branches: shadow-to-shadow-free (S2F) and shadow-free-to-shadow-free (F2F), both of which take a real-world image and a shadow mask as inputs. The input shadow mask is a binary map where 0 represents non-shadow (black) regions and 1 represents shadow (white) regions.

4.3.1 S2F and F2F sub-branches

In S2F, a shadow mask image $M^{'}$ consisting entirely of zeros and a real shadow image R_s are used as inputs to the generator G_s to produce an image $\widetilde{R}_s^{'}$ without additional shadows:

$$\widetilde{R}_s^{'} = G_s(R_s, M^{'}). \tag{5}$$

The generated shadow image $\widetilde{R}_s^{'}$ is then transformed into a shadow-free image $\widetilde{R}_n^{'}$ using the generator G_r:

$$\widetilde{R}_n^{'} = G_r(\widetilde{R}_s^{'}). \tag{6}$$

Subsequently, the discriminator D_r is employed to determine whether $\widetilde{R}_n^{'}$ is a real shadow-free image.

In F2F, a real shadow-free image R_n and a shadow mask $M^{''}$, which indicates the shadow region, are required as inputs. We use the method described in [28] to generate the shadow mask $M^{''}$. Then, with these inputs, generator G_s generates a shadow image $\widetilde{R}_s^{''}$ to deceive the discriminator D_s, making it difficult for D_s to distinguish whether it is a real shadow image:

$$\widetilde{R}_s^{''} = G_s(R_n, M^{''}). \tag{7}$$

The synthesized $\widetilde{R}_s^{''}$ is then used as input to generator G_r for shadow removal, producing a coarse shadow-free image $\widetilde{R}_n^{''}$.

$$\widetilde{R}_n^{''} = G_r(\widetilde{R}_s^{''}). \tag{8}$$

In this process, we find that the input to generator G_r is always the synthesized shadow image generated by G_s. To improve the shadow removal performance of G_r, we integrate MSP into G_r to reduce the impact of synthesized data in both S2F and F2F.

The architectures of G_s and G_r are identical, following the generator design proposed by Hu [15]. Each consists of three convolution layers with a stride of 2, followed by nine residual blocks for feature extraction, and finally three deconvolution layers to upsample the feature map. Instance normalization is applied after each convolution operation. For the discriminators D_r and D_s, we adopt the architecture proposed in PatchGAN [17].

4.3.2 Loss function

In S2F, we use identity loss to make the generated shadow image $\widetilde{R}_s^{'}$ close to the input shadow image R_s:

$$\mathcal{L}_{identity}(G_s) = \mathbb{E}_{R_s \sim p(R_s)} \left[\| G_s(R_s, M^{'}) - R_s \|_1 \right]. \quad (9)$$

For generator G_r and its discriminatorD_r, the objective function is optimized as follows:

$$\begin{aligned} \mathcal{L}_{GAN_r} &= \mathbb{E}_{R_n \sim p(R_n)} \left[\log(D_r(R_n)) \right] \\ &+ \mathbb{E}_{R_s \sim p(R_s)} \left[\log(1 - D_r(G_r(\widetilde{R}_s^{'}))) \right]. \end{aligned} \quad (10)$$

We incorporate the MSP into G_r. This process introduces a contrastive loss to help G_r better remove shadows. As shown in Fig. 3, the contrastive loss constraint between the output $\mathcal{S}(X, Y_s)$ of the last SFB and the semantic features extracted from input text T by clip text encoder C_{text} is defined as follows:

$$\mathcal{L}_{clip_{S2F}} = \frac{e^{cos(\mathcal{S}(X,Y_s), C_{text}(T))/\tau}}{e^{cos(\mathcal{S}(X,Y_s), C_{text}(T))/\tau} + e^{\frac{1}{\tau}}}, \quad (11)$$

where X is the output of the penultimate residual block in G_r, $Y_s = C_{image}(R_s)$,τ denotes the temperature parameter which we set to 0.5 in experiment.

In F2F, the adversarial loss for generator G_s and discriminator D_s is formulated as:

$$\begin{aligned} \mathcal{L}_{\text{GAN}_s} &= \mathbb{E}_{R_s \sim p(R_s)} \left[\log(D_s(R_s)) \right] \\ &+ \mathbb{E}_{R_n \sim p(R_n)} \left[\log(1 - D_s(G_s(R_n, M^{''}))) \right]. \end{aligned} \quad (12)$$

MSP is also used in F2F, so we define a loss similar to $\mathcal{L}_{clip_{S2F}}$ that makes the shadow removal result close to the input text T:

$$\mathcal{L}_{clip_{F2F}} = \frac{e^{cos(\mathcal{S}(X,Y_n), C_{text}(T))/\tau}}{e^{cos(\mathcal{S}(X,Y_n), C_{text}(T)))/\tau} + e^{\frac{1}{\tau}}}, \quad (13)$$

where $Y_n = C_{image}(R_n)$.

To ensure that $\widetilde{R}_n^{''}$ is similar to the original input real shadow-free image R_n, we use cycle consistency loss to optimize the mapping functions in G_s and G_r:

$$\mathcal{L}_{cycle}(G_s, G_r) = \mathbb{E}_{R_n \sim p(R_n)} \left[\| G_r(G_s(R_n, M^{''})) - R_n \|_1 \right]. \quad (14)$$

To further emphasize that the shadow region in $\widetilde{R}_n^{''}$ guided by the shadow mask matches the content in the input image R_n, we apply a shadow loss as follow:

$$\mathcal{L}_{shadow}(G_s, G_r) = \mathbb{E}_{R_n \sim p(R_n)} \left[\| M \odot (G_r(G_s(R_n M^{''})) - R_n) \|_1 \right]. \quad (15)$$

In summary, the total loss in SG-GAN is defined as:

$$\begin{aligned} \mathcal{L} = \lambda_1 \mathcal{L}_{\text{identity}} + \lambda_2 (\mathcal{L}_{\text{GAN}_s} + \mathcal{L}_{\text{GAN}_r}) \\ + \lambda_3 (\mathcal{L}_{clip_{S2F}} + \mathcal{L}_{clip_{F2F}}) + \lambda_4 (\mathcal{L}_{\text{cycle}} + \mathcal{L}_{\text{shadow}}), \end{aligned} \quad (16)$$

where λ_1, λ_2, λ_3, and λ_4 are weights balancing different loss terms. In our experiment, we set λ_1, λ_2, λ_3, and λ_4 to 5, 1, 0.5, and 10, respectively.

4.4. Diffusion-based restoration module

At this stage, we employ IR-SDE [33] as the diffusion framework for our model. This framework allows us to better understand and control the image generation process by simulating the image degradation process. The key idea behind our diffusion model is to combine a mean-reverting SDE with a maximum likelihood objective for neural network training, which naturally transforms high-quality images into degraded low-quality images, regardless of the complexity of the degradation.

Diffusion framework According to [33], the forward process of a mean-reverting SDE is defined as:

$$dx_t = \theta_t(\mu - x_t)dt + \sigma_t dw, \quad (17)$$

where θ_t and σ_t are time-varying positive parameters representing the mean reversion rate and stochastic volatility, respectively. Here, μ is the state mean, and w denotes Brownian motion.

In our model, a coarse shadow removal result $\widetilde{R}_n^{''}$ is obtained in the second branch of SG-GAN. We use $\widetilde{R}_n^{''}$ and the corresponding clean input image R_n to construct paired data (LQ and GT) for our DBRM. Given paired images, we set R_n as the initial state x_0 and $\widetilde{R}_n^{''}$ as μ, with a fixed noise level λ.

Ensure $\sigma_t^2/\theta_t = 2\lambda^2$ for all t, the forward process solution is:

$$x(t) = \widetilde{R}_n^{''} + (x(s) - \widetilde{R}_n^{''})e^{-\bar{\theta}s} + \int_s^t \sigma_z e^{-\bar{\theta}z} dw(z), \quad (18)$$

where $\bar{\theta}s = \int_s^t \theta_z dz$. The transition kernel is:

$$p(x(t) \mid x(s)) = \mathcal{N}(x(t) \mid m_s(x(s)), v_s), \quad (19)$$

where m_s is the mean and v_s is the variance. The forward SDE iteratively transforms the GT image R_n into the LQ image $\widetilde{R}_n^{''}$ with added noise:

$$p(x_t \mid x_{t-1}) = \mathcal{N}(x_t \mid m_{t-1}(x_{t-1}), v_{t-1}). \quad (20)$$

A notable property of this process is that noisy data x_t can be sampled from x_0 in closed form:

$$p_t(x) = \mathcal{N}\left(x(t) \mid m_t(x), v_t\right), \tag{21}$$

with $m_t = \widetilde{R}_n'' + (x(0) - \widetilde{R}_n'')e^{-\bar{\theta}t}$ and $v_t = \lambda^2\left(1 - e^{-2\bar{\theta}t}\right)$.

The reverse-time representation from [1] is:

$$dx = [\theta_t(\widetilde{R}_n'' - x) - \sigma_t^2 \nabla_x \log p_t(x)]dt + \sigma_t dw, \tag{22}$$

where $\nabla_x \log p_t(x)$ is the score of the marginal distribution at time t. Given the GT image R_n, we compute the score function as:

$$\nabla_x \log p_t(x) = -\frac{x(t) - m_t}{v_t}. \tag{23}$$

We then train a conditional time-dependent neural network $\tilde{\epsilon}_\phi(x_t, \widetilde{R}_n'', t)$ to estimate the noise. Sampling x_t is done according to $x_t = m_t(x) + \sqrt{v_t}\ \epsilon_t$, where $\epsilon_t \sim \mathcal{N}(0, I)$. The score can then be directly computed from the noise:

$$\nabla_x \log p_t(x) = -\frac{\epsilon_t}{\sqrt{v_t}}. \tag{24}$$

4.4.1 Network architecture

As shown in Fig. 3, our noise prediction network is based on the Nonlinear Activated Free (NAF) block [3]. The NAF block replaces traditional nonlinear activations (such as ReLU and GELU) with a SimpleGate unit. Given an input, SimpleGate splits it into two features along the channel dimension and then uses a linear gate to compute the output. The SimpleGate unit is added after the depth-wise convolution and between the two fully connected layers. Additionally, we introduce multi-layer perceptual processing and time embedding for each NAF block.

Loss function An alternative maximum likelihood objective aims to find the optimal trajectory $x_{1:T}$ given the high-quality image x_0, stabilizing training and recovering more accurate images. Following IR-SDE [33], we train our prediction network with a maximum likelihood loss which specifies the optimal reverse path x_{t-1}^* for all times:

$$\begin{aligned} x_{t-1}^* = &\frac{1 - e^{-2\bar{\theta}_{t-1}}}{1 - e^{-2\bar{\theta}_t}} e^{-\theta_t'}(x_t - \widetilde{R}_n'') \\ &+ \frac{1 - e^{-2\theta_t'}}{1 - e^{-2\bar{\theta}_t}} e^{-\bar{\theta}_{t-1}}(R_n - \widetilde{R}_n'') + \widetilde{R}_n'', \end{aligned} \tag{25}$$

where $\theta_i' = \int_{i-1}^{i} \theta_t\, dt$. Then, we choose to optimize the noise network $\tilde{\epsilon}_\phi$ to make IR-SDE reverse as the optimal trajectory, as

$$\mathcal{L}_{diff} = \sum_{t=1}^{T} \gamma_t \mathbb{E}\left[\left\|x_t - (dx_t)\tilde{\epsilon}_\phi - x_{t-1}^*\right\|\right], \tag{26}$$

where $\gamma_1, ..., \gamma_T$ are positive weights and $\{x_t\}_{t=0}^T$ denotes the discretization of the diffusion process. $(dx)\tilde{\epsilon}_\phi$ denotes the reverse-time SDE in Eq.17 and its score is predicted by the noise network $\tilde{\epsilon}_\phi$. $x_t - (dx_t)\tilde{\epsilon}_\phi$ is the reverse x_{t-1}. Once trained, we can use the network $\tilde{\epsilon}_\phi$ to generate high-quality images by sampling a noisy state x_T and iteratively solving the Eq.17 with a numerical scheme.

5. Experiment

5.1. Datasets and evaluation metrics

Dataset We utilize four state-of-the-art shadow removal datasets: ISTD [43], AISTD [21], SRD [36], and USR [15]. ISTD comprises 1,870 triplets of shadow images, shadow-free images, and shadow masks, with 1,330 triplets for training and 540 triplets for testing. AISTD is an adjusted dataset that further corrects the color inconsistency problem of images from ISTD. The SRD dataset consists of 2,680 training pairs and 408 testing pairs of shadow and shadow-free images. We use the predicted masks provided by DHAN [5] for training and testing. USR is an unpaired shadow removal dataset with 2,445 shadow images and 1,770 shadow-free images. The shadow images are divided into 1,956 images for training and 489 for testing, while all 1,770 shadow-free images are used for training.

Evaluation metrics In our experiments, we follow [23] calculate Root-Mean-Square Error (RMSE) in the LAB color space, and employ the Structure Similarity (SSIM), Peak Signal-to-Noise ratio (PSNR), and Learned Perceptual Image Patch Similarity (LPIPS) as evaluation metrics for comparisons. Generally, higher PSNR and SSIM values are preferred, while lower RMSE and LPIPS values indicate better performance. We provide the metrics measured on the shadow region, non-shadow region and whole image for reference. Since the USR dataset is unpaired, we use the Fréchet Inception Distance (FID) and Kernel Inception Distance (KID) to evaluate the results.

5.2. Experimental settings

We implement our methods using PyTorch [35] and a single NVIDIA GeForce RTX 3090 GPU. At first stage, we initialise our SG-GAN using a Gaussian distribution with a mean of 0 and a standard deviation of 0.02. We employ the Adam optimiser to train our network with the first and the second momentum setting to 0.5 and 0.999, respectively. We train the whole model for 200 epochs and the base learning rate is set to 2×10^{-4} for the first 100 epochs and then we apply a linear decay strategy to decrease it to 0 for the rest epochs. Additionally, horizontal flipping and random cropping strategy purposed in [28] is applied to the training data for data augmentation. The network training involves both two branches, and they impact each other. At second stage, for training the diffusion model, we fix the noise level at 50

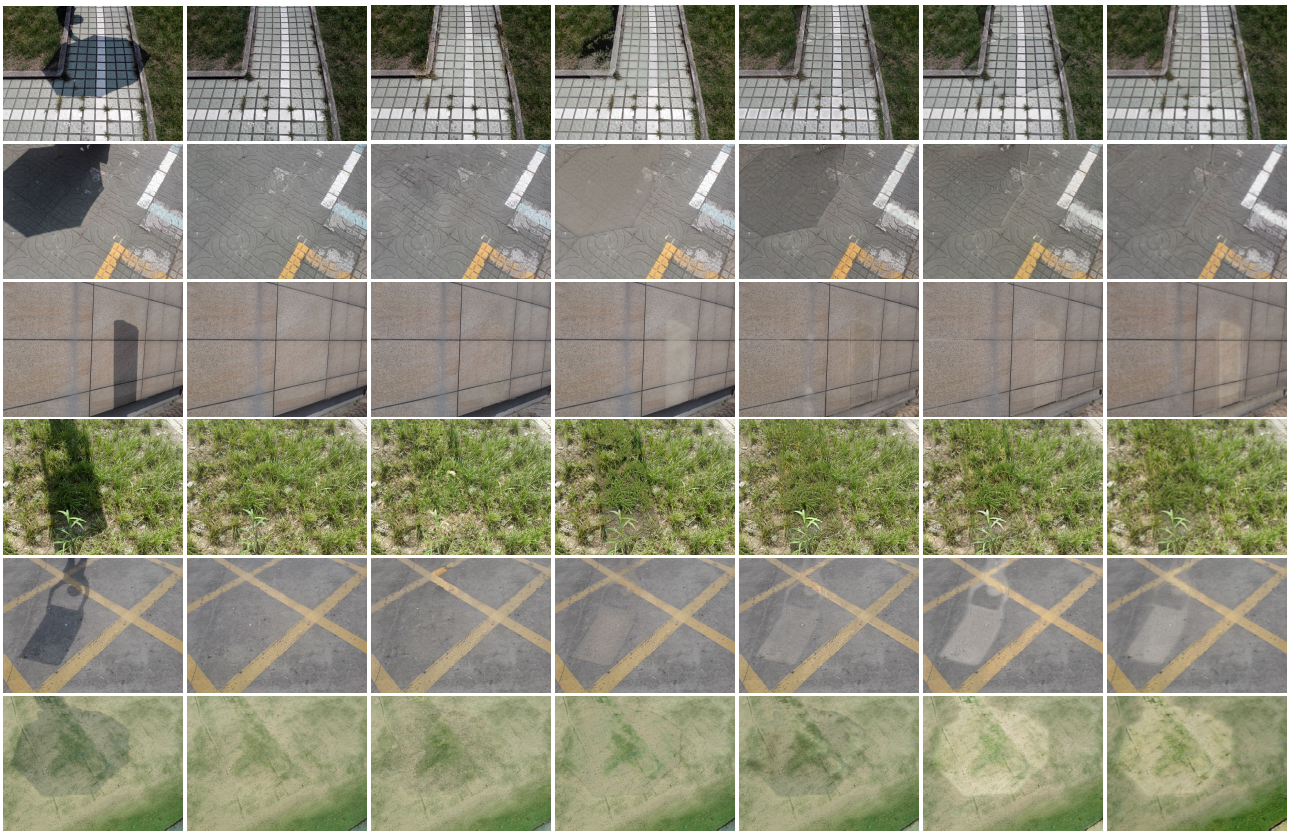


Figure 5. Visualisation comparisons results on six real-world challenging samples from the ISTD (rows 1-3) and AISTD (rows 4-6) datasets.

Scheme	Methods	Shadow Region			Non-shadow Region			All			
		RMSE↓	PSNR↑	SSIM↑	RMSE↓	PSNR↑	SSIM↑	RMSE↓	PSNR↑	SSIM↑	LPIPS↓
Supervised	DHAN	7.53	35.82	0.989	5.33	30.95	0.971	5.68	29.09	0.953	**0.027**
	Sp+M-Net	7.13	35.08	0.984	**3.16**	**36.38**	**0.979**	**3.92**	31.89	0.953	0.079
	AEFNet	7.91	34.71	0.975	5.51	28.61	0.880	5.88	27.19	0.945	0.045
	ShadowDiffusion	**4.10**	**40.06**	**0.996**	4.18	33.00	0.973	4.16	**32.20**	**0.967**	-
Unsupervised	Mask-ShadowGAN	13.00	30.53	0.977	6.07	28.86	0.960	6.96	25.72	0.925	0.064
	DC-ShadowNet	11.89	31.27	0.966	7.84	27.20	0.910	7.03	25.51	0.865	0.104
	LG-ShadowNet	10.92	31.23	**0.978**	6.30	27.67	0.967	6.29	26.39	0.935	0.058
	G2R-ShadowNet	10.53	32.32	0.975	7.09	27.32	**0.976**	7.33	25.70	**0.941**	0.047
	S3R-Net	12.16	-	-	6.38	-	-	7.12	-	-	-
	Ours	**10.00**	**32.68**	0.969	**5.54**	**30.96**	0.970	**6.26**	**27.94**	0.930	**0.038**

Table 1. Quantitative comparison results of our methods with the state-of-the-art methods on ISTD 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.

and set the number of diffusion denoising steps to 100. The batch sizes are set to 8 and the training patches are 256×256 pixels. We use the Lion optimizer [4] with β_1 = 0.9 and β_2 = 0.99. The initial learning rate is set to 3×10^{-5} and decayed to 1e-7 by the Cosine scheduler. The noise level is fixed to 50 and the number of diffusion denoising steps is set to 100. We train DBRM for 400 000 iterations, which takes for about 4 days on the GPU.

5.3. Comparison with the state-of-the-art on paired datasets

In this subsection, we compare our full model on the ISTD, AISTD and SRD datasets with several state-of-the-art methods, including supervised methods which are trained with paired shadow and shadow-free images: DHAN [5], SP+M-Net [21], AEFNet [9] and ShadowDiffusion [10], DSC [14]; unsupervised methods training without

Figure 6. Visualisation comparisons results on five real-world challenging samples from the SRD dataset.

Scheme	Methods	Shadow Region			Non-shadow Region			All			
		RMSE↓	PSNR↑	SSIM↑	RMSE↓	PSNR↑	SSIM↑	RMSE↓	PSNR↑	SSIM↑	LPIPS↓
Supervised	DHAN	9.57	32.92	0.987	7.41	27.15	0.972	7.77	25.66	0.954	**0.026**
	Sp+M-Net	5.91	37.60	0.990	2.99	36.02	0.976	3.46	32.94	0.962	0.085
	AEFNet	6.55	36.04	0.978	3.77	31.16	0.892	4.22	29.45	0.861	0.046
	ShadowDiffusion	**4.60**	**40.13**	**0.997**	**2.74**	**36.48**	**0.979**	**2.91**	**35.66**	**0.974**	-
Unsupervised	Mask-ShadowGAN	11.28	31.50	0.981	3.90	32.63	0.967	4.97	28.11	0.936	0.063
	DC-ShadowNet	10.81	32.15	0.978	3.46	35.50	0.974	4.61	29.09	0.940	0.051
	LG-ShadowNet	9.90	32.42	**0.982**	3.18	34.01	0.976	4.25	29.31	**0.947**	0.049
	S3R-Net	12.86	-	-	4.43	-	-	5.71	-	-	-
	Ours	**9.48**	**33.63**	0.972	**3.09**	**36.01**	**0.978**	**4.06**	**31.09**	0.940	**0.034**

Table 2. Quantitative comparison results of our methods with the state-of-the-art methods on AISTD dataset.

paired shadow and shadow-free images: G2R-ShadowNet [28], Mask-ShadowGAN [15], DC-ShadowNet [18], LG-ShadowNet [27] and S3R-Net [20]. All of the shadow removal results by the competing methods are quoted from the original papers or reproduced using their official implementations.

Table 1 shows the quantitative results on the ISTD dataset. The supervised methods share the same type of training data, including shadow and shadow-free image pairs. They learn the mapping from shadow images to shadow-free images based on training pairs. Our method achieves results in shadow-free region and whole image that are comparable to other deep neural networks trained on paired images, and in some metrics, it even surpasses some supervised methods. For instance, LPIPS metric of our results for the whole image is better than those of Sp+M-net and AEFNet. Moreover, several unsupervised methods, such as Mask-ShadowGAN, LG-ShadowNet, DC-ShadowNet, G2R-ShadowNet and S3R-Net, we can see that our method significantly outperforms these three methods. Our method outperforms LG-ShadowNet which is suboptimal on most metrics, especially the results for non-shadow region and the whole image improve by about 3.3dB and 1.6dB in PSNR, respectively, except SSIM value is slightly lower. Additionally, our LPIPS, PSNR and RMSE are clearly better than all the unsupervised methods in the table. When compared to the state-of-the-art unsupervised method S3R-Net, our method performs better on RMSE. Table 2 shows the quantitative results on the testing sets over AISTD. It is clear that our method outperform su-

Scheme	Methods	Shadow Region			Non-shadow Region			All			
		RMSE↓	PSNR↑	SSIM↑	RMSE↓	PSNR↑	SSIM↑	RMSE↓	PSNR↑	SSIM↑	LPIPS↓
Supervised	DHAN	8.94	33.67	0.978	4.80	34.79	0.979	5.67	30.51	0.949	-
	DSC	8.62	30.65	0.960	4.41	31.94	0.965	5.71	27.76	0.903	0.150
	AEFNet	6.55	36.04	0.978	3.77	31.16	0.892	4.22	29.45	0.861	**0.102**
	ShadowDiffusion	**4.98**	**38.72**	**0.987**	**3.44**	**37.78**	**0.985**	**3.63**	**34.73**	**0.970**	-
Unsupervised	Mask-ShadowGAN	11.90	29.71	0.960	5.87	31.77	0.968	7.66	26.96	0.915	0.096
	DC-ShadowNet	10.81	30.15	0.970	4.96	33.50	0.974	6.21	28.67	0.930	0.087
	LG-ShadowNet	9.90	30.42	**0.972**	4.71	32.01	**0.976**	5.88	28.31	**0.937**	0.103
	G2R-ShadowNet	16.05	25.80	0.925	5.34	31.17	0.972	9.03	24.17	0.878	0.601
	Ours	**9.20**	**31.20**	0.963	**4.18**	**33.94**	0.974	**5.45**	**29.49**	0.931	**0.059**

Table 3. Quantitative comparison results of our methods with the state-of-the-art methods on SRD dataset.

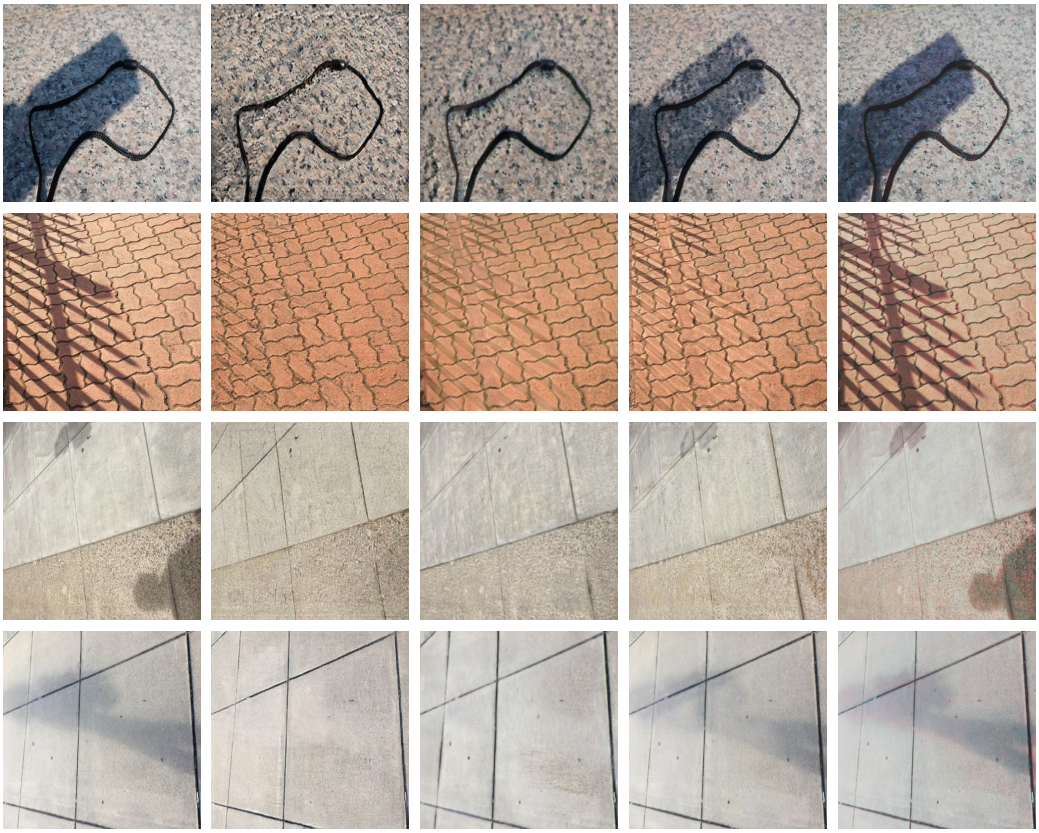


Figure 7. Visualisation comparisons results on five real-world challenging samples from the USR dataset.

pervised methods DHAN and AEFNet in the non-shadow region and the whole image. It improves the PSNR from 29.31dB to 31.09dB, compared to the suboptimal method LG-ShadowNet. Fig. 5 shows the qualitative results of our method and other state-of-art methods on six challenging sample images in the ISTD(row 1-3) and AISTD (rows 4-6) datasets. It is worth noting that since we are unable to obtain the visualization results of S3R-Net, no qualitative comparison is made with S3R in Fig. 5. Compared with other methods, our method can produce more realistic results with less artifacts and better preserve the texture details occluded by shadows. Moreover, the color in the shadow region is more consistent with the surrounding area using our method.

Next, we compare our method with the state-of-the-art methods on the SRD dataset, with quantitative results presented in Table 3. Among the unsupervised methods, just like before our method performs the best on SRD, followed by LG-ShadowNet [27]. On this dataset, we outperform other unsupervised methods in nearly all metrics, except for SSIM. In Fig. 6, the outputs generated by our competitors exhibit sharp shadow edges, whereas our results have significantly smoother shadow boundaries. Additionally, our

Input w/o S2F w/o MSP w/o DBRM Ours Ground truth

Figure 8. Visual comparisons result of ablation study on the use of each component for our method.

Metric	Mask-ShadowGAN	LG-ShadowNet	DC-ShadowNet	Ours
FID↓	285.03	225.01	220.02	**204.02**
KID↓	0.060	0.016	0.020	**0.014**

Table 4. Quantitative comparison results of our methods with the state-of-the-art methods on USR testing set.

Methods	RMSE↓	PSNR↑	SSIM↑	LPIPS↓
w/o DBRM	7.27	25.55	**0.935**	0.054
SG-GAN w/o S2F	10.36	21.45	0.892	0.122
SG-GAN w/o MSP	7.99	24.89	0.902	0.065
Ours	**6.26**	**27.94**	0.930	**0.038**

Table 5. Ablation study on the choices of different component for our method on ISTD testing set.

Methods	RMSE↓	PSNR↑	SSIM↑	LPIPS↓
w/o $\mathcal{L}_{GAN_s}$	8.75	24.34	0.838	0.119
w/o $\mathcal{L}_{GAN_r}$	10.05	21.64	0.900	0.115
w/o $\mathcal{L}_{cycle}$	8.40	24.58	0.825	0.126
w/o $\mathcal{L}_{shadow}$	8.03	24.79	0.877	0.111
w/o $\mathcal{L}_{identity}$	13.54	19.39	0.845	0.182
w/o $\mathcal{L}_{clip}$	7.71	25.03	0.907	0.088
SG-GAN	**7.27**	**25.55**	**0.935**	**0.054**

Table 6. Ablation study on the choices of the loss functions for our SG-GAN on ISTD testing set.

samples show a noticeably sharper appearance in shadow region. We believe we have demonstrated that our approach can produce the most visually pleasing results to the human eye.

5.4. Comparison with the state-of-the-art on unpaired datasets

We compare our method with several unsupervised ones, including Mask-ShadowGAN, LG-ShadowNet, and DC-ShadowNet, on the unpaired dataset USR. We employee FID and KID indicators to quantitatively analyze the outcomes. Quantitative results are shown in Table 4 , where it can be seen that our method outperforms the other comparison methods on both metrics. Especially on the FID metric, our method outperforms the second-best DC-Shadow method by nearly 16 points. A visual comparison of real-world samples, as depicted in Fig. 7, also indicates that our approach performs outstandingly for complex shadows (rows 1-2), multiple shadows (row 4), and subtle shadows (row 5).

5.5. Ablation study

To validate the efficacy of each pivotal component within our proposed method, we trained and evaluated several model variants on the ISTD dataset. First, we propose to utilize the DBRM to suppress the artifacts of shadow removal results. So we validate DBRM by removing it from our complete model, retaining only SG-GAN. Additionally, we train SG-GAN without the S2F and MSP components to evaluate their individual contributions. The quantitative results are reported in Table 5. Subsequently, we conduct ablation studies on loss functions we proposed. These studies involved training SG-GAN without specific loss terms to demonstrate the effectiveness of each loss function. The quantitative results are reported in Table 6.

As show in Table 5, we observe that performances of our method reduces across all metrics except SSIM when DBRM is omitted. Comparing row 2 to row 4, we find that the S2F branch is crucial, providing substantial performance gains in terms of all the metrics. Then, when MSP is excluded, there is also a certain drop in performance. In Table 6, we find that $\mathcal{L}_{GAN_s}$ is important and brings performance improvement. Rows 2 and 5 show a significant decline in SG-GAN's performance without $\mathcal{L}_{GAN_r}$ and $\mathcal{L}_{identity}$, especially for the $\mathcal{L}_{identity}$. Then, when $\mathcal{L}_{clip}$, $\mathcal{L}_{cycle}$, and $\mathcal{L}_{shadow}$ is respectively removed from the total loss, the performance of SG-GAN shows a slight decrease. As shown in Figure 8 and Figure 9, the qualitative results are generally consistent with the aforementioned quantitative results in demonstrating the effectiveness of each component. Compared to the model trained with all components, other variants trained with subsets of these components may exhibit noticeable artifacts in the results. For instance, the result of SG-GAN w/o $\mathcal{L}_{shadow}$ in Figure 9 and the result of w/o S2F in Figure 8, artifact in the shadow region is very prominent.

5.6. Effectiveness of general-purpose MSP

We propose a general MSP to mitigate the impact of synthetic images on the performance of the shadow removal

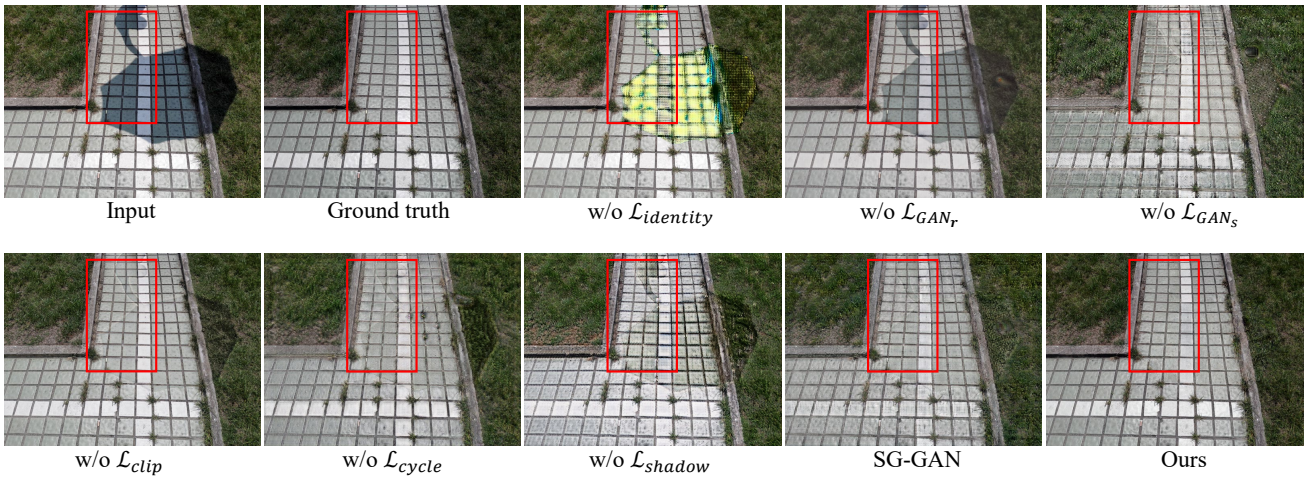
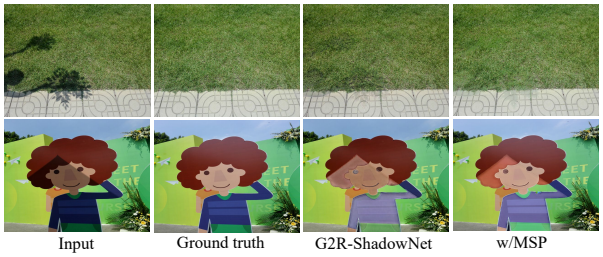
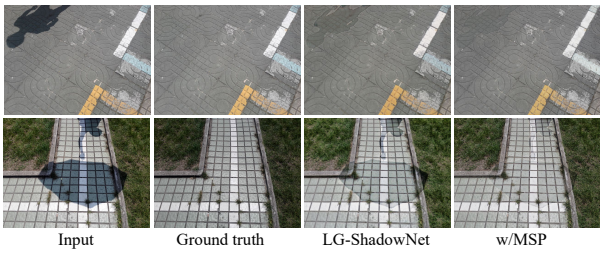


Figure 9. Visual comparisons result for ablation study on the use of each loss term in SG-GAN.



(a) Comparison results on G2R-ShadowNet



(b) Comparison results on LG-ShadowNet

(c) Comparison results on Mask-ShadowGAN

Figure 10. Visualisation comparisons results on the effectiveness of the MSP.

network. We experimentally validate the effectiveness of MSP on the ISTD dataset. We conduct experiments not only on our own model but also by integrating our proposed MSP into three unsupervised methods, G2R-ShadowNet [28], Mask-ShadowGAN[15] and LG-ShadowNet [27], which require synthetic shadows for shadow removal network training.

Methods	RMSE↓	PSNR↑	SSIM↑	LPIPS↓
Ours w/o MSP	6.47	27.07	0.915	0.048
Ours	**6.26**	**27.94**	**0.930**	**0.038**
Mask-ShadowGAN	6.96	25.72	0.925	0.064
Mask-ShadowGAN w/ MSP	**6.53**	**26.39**	**0.935**	**0.053**
G2R-ShadowNet	7.33	25.70	0.941	0.047
G2R-ShadowNet w/ MSP	**6.83**	**26.57**	**0.958**	**0.033**
LG-ShadowNet	6.29	26.39	0.935	0.058
LG-ShadowNet w/ MSP	**6.10**	**26.80**	**0.943**	**0.051**

Table 7. Ablation studies on the effectiveness of the MSP on the ISTD testing set.

Quantitative results are shown in Table 7. The results clearly demonstrate that the performance of our method and the other two methods has shown varying degrees of improvement after integrating MSP, particularly in the case of G2R-ShadowNet [28]. Although the numerical improvements in the metrics appear to be minor, the corresponding visual results in Figure 10 demonstrate a significant enhancement in shadow removal performance when integrating our MSP module into the original method. The color consistency in the recovered shadow regions is noticeably improved, rendering the output closer to a realistic shadow-free image. It is noteworthy that both our method and Mask-ShadowGAN [15] are trained in RGB space, whereas G2R-ShadowNet [28] is trained in LAB space and LG-ShadowNet is trained in both LAB and RGB space. This outcome indicates the universal applicability of our MSP across different color spaces.

6. Conclusion

In this paper, we propose a novel coarse-to-fine framework for self-supervised shadow removal, comprising two stages: a coarse processing stage and a refined restoration stage, implemented through the SG-GAN and DBRM networks, respectively. In SG-GAN, shadows are first generated and then removed, creating paired training data for refinement in DBRM. Additionally, we design a general-purpose Multi-modal Semantic Prompter module to mitigate the impact of synthetic data on network performance. The coarse results are further refined by DBRM's powerful generative capabilities, restoring texture details and resolving edge artifacts in shadowed regions. Extensive experiments demonstrate the effectiveness of our approach, showing competitive performance on the ISTD, AISTD, SRD and USR datasets compared to other state-of-the-art techniques.

References

[1] B. D. Anderson. Reverse-time diffusion equation models. *Stochastic Processes and their Applications*, 12(3):313–326, 1982. 3, 7

[2] E. Arbel and H. Hel-Or. Shadow removal using intensity surfaces and texture anchor points. *IEEE transactions on pattern analysis and machine intelligence*, 33(6):1202–1216, 2010. 1

[3] L. Chen, X. Chu, X. Zhang, and J. Sun. Simple baselines for image restoration. In *European conference on computer vision*, pages 17–33. Springer, 2022. 7

[4] X. Chen, C. Liang, D. Huang, E. Real, K. Wang, H. Pham, X. Dong, T. Luong, C.-J. Hsieh, Y. Lu, et al. Symbolic discovery of optimization algorithms. *Advances in Neural Information Processing Systems*, 36, 2024. 8

[5] X. Cun, C.-M. Pun, and C. 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. 1, 7, 8

[6] V. De Bortoli, E. Mathieu, M. Hutchinson, J. Thornton, Y. W. Teh, and A. Doucet. Riemannian score-based generative modelling. *Advances in Neural Information Processing Systems*, 35:2406–2422, 2022. 3

[7] L. Fang and F. Yu. Moving object detection algorithm based on removed ghost and shadow visual background extractor. *Laser & Optoelectronics Progress*, 56(13):131002, 2019. 1

[8] G. D. Finlayson, S. D. Hordley, C. Lu, and M. S. Drew. On the removal of shadows from images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(1):59–68, 2006. 1, 3

[9] L. Fu, C. Zhou, Q. Guo, F. Juefei-Xu, H. Yu, W. Feng, Y. Liu, and S. Wang. Auto-exposure fusion for single-image shadow removal. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10571–10580, 2021. 8

[10] L. Guo, C. Wang, W. Yang, S. Huang, Y. Wang, H. Pfister, and B. Wen. Shadowdiffusion: When degradation prior meets diffusion model for shadow removal. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14049–14058. IEEE, 2023. 2, 8

[11] L. Guo, C. Wang, W. Yang, Y. Wang, and B. Wen. Boundary-aware divide and conquer: A diffusion-based solution for unsupervised shadow removal. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13045–13054, 2023. 3

[12] R. Guo, Q. Dai, and D. Hoiem. Paired regions for shadow detection and removal. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(12):2956–2967, 2013. 1, 3

[13] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 2, 3

[14] X. Hu, C.-W. Fu, L. Zhu, J. Qin, and P.-A. Heng. Direction-aware spatial context features for shadow detection and removal. *IEEE transactions on pattern analysis and machine intelligence*, 42(11):2795–2808, 2019. 8

[15] X. Hu, Y. Jiang, C. W. Fu, and P. A. Heng. Mask-shadowgan: Learning to remove shadows from unpaired data. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 1, 3, 4, 6, 7, 9, 12

[16] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, and W. Liu. Ccnet: Criss-cross attention for semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 603–612, 2019. 5

[17] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017. 6

[18] Y. Jin, A. Sharma, and R. 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. 9

[19] B. Kim, Y. Oh, and J. C. Ye. Diffusion adversarial representation learning for self-supervised vessel segmentation. *arXiv preprint arXiv:2209.14566*, 2022. 3

[20] N. Kubiak, A. Mustafa, G. Phillipson, S. Jolly, and S. Hadfield. S3r-net: A single-stage approach to self-supervised shadow removal. *arXiv preprint arXiv:2404.12103*, 2024. 9

[21] H. Le and D. Samaras. Shadow removal via shadow image decomposition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8578–8587, 2019. 7, 8

[22] H. Le and D. Samaras. From shadow segmentation to shadow removal. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, pages 264–281. Springer, 2020. 2

[23] H. Le and D. Samaras. Physics-based shadow image decomposition for shadow removal. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12):9088–9101, 2021. 7

[24] X. Li, Y. Ren, X. Jin, C. Lan, X. Wang, W. Zeng, X. Wang, and Z. Chen. Diffusion models for image restoration and enhancement–a comprehensive survey. *arXiv preprint arXiv:2308.09388*, 2023. 2

[25] Z. Li and N. Snavely. Learning intrinsic image decomposition from watching the world. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9039–9048, 2018. 1

[26] Y. Liu, Z. Ke, K. Xu, F. Liu, Z. Wang, and R. W. Lau. Recasting regional lighting for shadow removal. *arXiv e-prints*, pages arXiv–2402, 2024. 1, 2, 3

[27] Z. Liu, H. Yin, Y. Mi, M. Pu, and S. Wang. Shadow removal by a lightness-guided network with training on unpaired data. *IEEE Transactions on Image Processing*, 30:1853–1865, 2021. 3, 9, 10, 12

[28] Z. Liu, H. Yin, X. Wu, Z. Wu, Y. Mi, and S. Wang. From shadow generation to shadow removal. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4927–4936, 2021. 1, 3, 4, 5, 7, 9, 12

[29] C. Lu, Y. Zhou, F. Bao, J. Chen, C. Li, and J. Zhu. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. *Advances in Neural Information Processing Systems*, 35:5775–5787, 2022. 3

[30] S. Lu, Y. Liu, and A. W.-K. Kong. Tf-icon: Diffusion-based training-free cross-domain image composition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2294–2305, 2023. 2

[31] S. Lu, Z. Wang, L. Li, Y. Liu, and A. W.-K. Kong. Mace: Mass concept erasure in diffusion models. *arXiv preprint arXiv:2403.06135*, 2024. 2

[32] A. Lugmayr, M. Danelljan, A. Romero, F. Yu, R. Timofte, and L. 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. 3

[33] Z. Luo, F. K. Gustafsson, Z. Zhao, J. Sjolund, and T. B. Schon. Image restoration with mean-reverting stochastic differential equations, 2023. 6, 7

[34] K. Mei, L. Figueroa, Z. Lin, Z. Ding, S. Cohen, and V. M. Patel. Latent feature-guided diffusion models for shadow removal. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4313–4322, 2024. 3

[35] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019. 7

[36] L. Qu, J. Tian, S. He, Y. Tang, and R. W. H. Lau. Deshadownet: A multi-context embedding deep network for shadow removal. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 3, 7

[37] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 4

[38] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. 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

[39] C. Saharia, W. Chan, H. Chang, C. Lee, J. Ho, T. Salimans, D. Fleet, and M. Norouzi. Palette: Image-to-image diffusion models. In *ACM SIGGRAPH 2022 conference proceedings*, pages 1–10, 2022. 2

[40] C. Saharia, J. Ho, W. Chan, T. Salimans, D. J. Fleet, and M. Norouzi. Image super-resolution via iterative refinement. *IEEE transactions on pattern analysis and machine intelligence*, 45(4):4713–4726, 2022. 3

[41] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR, 2015. 3

[42] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020. 3

[43] J. Wang, X. Li, and J. Yang. Stacked conditional generative adversarial networks for jointly learning shadow detection and shadow removal. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018. 3, 7

[44] C. Xiao, R. She, D. Xiao, and K. L. Ma. Fast shadow removal using adaptive multi-scale illumination transfer. *Computer Graphics Forum*, 32(8):207–218, 2013. 1, 3

[45] Z. Xiao, K. Kreis, and A. Vahdat. Tackling the generative learning trilemma with denoising diffusion gans. *arXiv preprint arXiv:2112.07804*, 2021. 2

[46] K. Xu, G. P. Hancke, and R. W. Lau. Learning image harmonization in the linear color space. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12570–12579, 2023. 3

[47] L. Zhang, C. Long, Q. Yan, X. Zhang, and C. Xiao. Cla-gan: A context and lightness aware generative adversarial network for shadow removal. *Computer Graphics Forum*, 2020. 1

[48] C. Zhao, W. Cai, C. Dong, and C. Hu. Wavelet-based fourier information interaction with frequency diffusion adjustment for underwater image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2024. 3

[49] C. Zhao, W. Cai, C. Hu, and Z. Yuan. Cycle contrastive adversarial learning with structural consistency for unsupervised high-quality image deraining transformer. *Neural Networks*, page 106428, 2024. 2

[50] D. Zhou, Y. Li, F. Ma, Z. Yang, and Y. Yang. Migc: Multi-instance generation controller for text-to-image synthesis. *arXiv preprint arXiv:2402.05408*, 2024. 2

[51] D. Zhou, Z. Yang, and Y. Yang. Pyramid diffusion models for low-light image enhancement. *arXiv preprint arXiv:2305.10028*, 2023. 2

[52] Y. Zhu, Z. Xiao, Y. Fang, X. Fu, Z. Xiong, and Z.-J. Zha. Efficient model-driven network for shadow removal. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 3635–3643, 2022. 1