Detail-Preserving Latent Diffusion for Stable Shadow Removal

Jiamin Xu
Hangzhou Dianzi University
superxjm@yeah.net

Yuxin Zheng
Hangzhou Dianzi University
yuxin6@hdu.edu.cn

Zelong Li
Hangzhou Dianzi University
jokerli@hdu.edu.cn

Chi Wang
Zhejiang University
wangchi1995@zju.edu.cn

Renshu Gu,
Hangzhou Dianzi University
renshugu@hdu.edu.cn

Weiwei Xu
Zhejiang University
xww@cad.zju.edu.cn

Gang Xu
Hangzhou Dianzi University
gxu@hdu.edu.cn

Abstract

Achieving high-quality shadow removal with strong generalizability is challenging in scenes with complex global illumination. Due to the limited diversity in shadow removal datasets, current methods are prone to overfitting training data, often leading to reduced performance on unseen cases. To address this, we leverage the rich visual priors of a pre-trained Stable Diffusion (SD) model and propose a two-stage fine-tuning pipeline to adapt the SD model for stable and efficient shadow removal. In the first stage, we fix the VAE and fine-tune the denoiser in latent space, which yields substantial shadow removal but may lose some high-frequency details. To resolve this, we introduce a second stage, called the detail injection stage. This stage selectively extracts features from the VAE encoder to modulate the decoder, injecting fine details into the final results. Experimental results show that our method outperforms state-of-the-art shadow removal techniques. The cross-dataset evaluation further demonstrates that our method generalizes effectively to unseen data, enhancing the applicability of shadow removal methods.

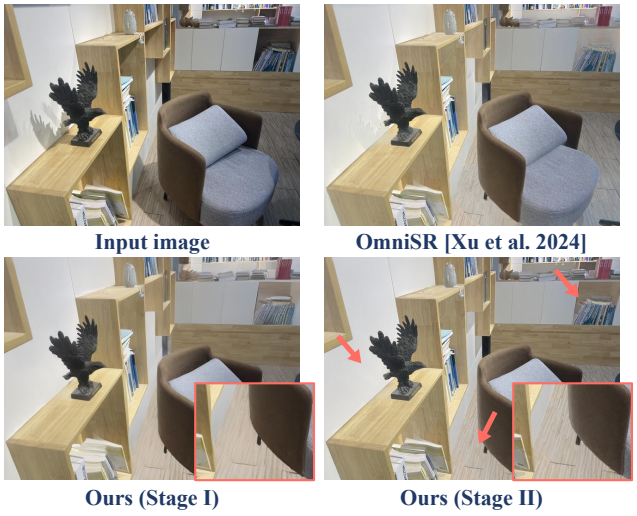

Figure 1. *Top*: For complex shadows in indoor scenes, current methods struggle to completely remove shadows, such as those behind the eagle sculpture. *Bottom*: Our method effectively removes the shadows, and with our second stage, the details from the original input, such as the wooden texture shown in the cropped image, are preserved.

1. Introduction

Shadows are natural phenomena that occur when an obstacle blocks global light transport. They provide crucial cues for interpreting the three-dimensional structure of objects and scenes [44]. However, shadows can also hinder computer vision tasks, such as object segmentation [8], tracking [32], and intrinsic decomposition [16, 17, 25, 43]. In recent years, deep-learning-based methods for shadow removal [4, 5, 24, 40] have shown significant advancements by learning pixel-wise mappings between shadowed images and their corresponding ground-truth shadow-free versions in a fully supervised manner. However, such approaches often risk overfitting the limited training datasets [14, 28, 35], leading to inadequate generalization.

Drawing inspiration from recent advancements in zero-shot dense prediction [12, 42], we leverage the robust and comprehensive visual priors of the pre-trained Stable Diffusion (SD) model [29], which has been trained on bil-

lions of images, to enhance the generalization capability of shadow removal. However, developing a fine-tuning strategy that effectively removes shadows while preserving clear details in both shadowed and non-shadowed areas is not trivial. Unlike low-frequency depth and normal prediction tasks [12, 42], applying the SD model to shadow removal can result in detail loss and misalignment between the conditional image and the shadow-free output (Fig. 1). It is because the SD model, which consists of a Variational Autoencoder (VAE) and a Latent Diffusion Model (LDM), is designed for text-to-image generation and can lead to lossy compression during the mapping from pixel space ($W \times H \times 3$) to latent space ($\frac{W}{8} \times \frac{H}{8} \times 4$).

To achieve stable and detail-preserving shadow removal using the pre-trained SD priors, we introduce a two-stage pipeline: 1) eliminating shadows in latent space by fixing the pre-trained VAE and fine-tuning the LDM and 2) injecting shadow-aware multi-resolution details from the original input into the VAE decoding process.

In the first stage, our method retains the pre-trained VAE and fine-tunes the LDM by conditioning it on the latent features of the input shadow image. We do not fine-tune the VAE, as it effectively handles shadow-free images despite not being specifically trained for this purpose, and the scale of our shadow-free images is limited. Our findings indicate that the latent space from the VAE, combined with the generative priors from the LDM, enhances shadow removal tasks, particularly when annotated training data is scarce. By operating in a low-resolution latent space rather than high-resolution pixel space, our method performs global self-attention with acceptable memory consumption, capturing more global contextual information compared to convolution and window-based self-attention [21].

The fine-tuned SD model in the first stage produces visually appealing results, but it may lose some local details. To further enhance the quality of shadow removal, we propose a second stage for injecting shadow-free details. In this stage, we introduce a Detail Injection (DI) module that enhances the decoded latent features from the pre-trained VAE decoder by incorporating fine details from the input images. The VAE parameters are kept fixed. To prevent shadows from being reintroduced as "details", our DI module is equipped with shadow perception capabilities. Inspired by the observation that the shadow mask in the current shadow removal dataset is derived from the binary difference between shadowed and shadow-free images, we utilize concatenated features from the input shadow images and the coarse, shadow-free images decoded from the LDM output as cues for shadow-free detail injection. By combining features from both image types, the DI module can implicitly detect shadows and learn to inject shadow-free details into the VAE decoders.

Note that Refusion [23] also performs latent-space diffusion for shadow removal. It captures input image information in the decoder [23] using a U-Net architecture specifically designed for large image shadow removal. However, the primary difference lies in the network design and fine-tuning strategy. Our method is designed to leverage the high-quality priors of the pre-trained VAE and LDM in the SD model, resulting in superior shadow removal quality. Our method can also handle high-resolution images (e.g., 1920×1440 in the WRSD dataset [35]), providing greater shadow removal efficiency compared to patch-based diffusion methods [22].

Overall, our contributions can be summarized as follows.

- We propose a two-stage fine-tuning pipeline to transform a pre-trained Stable Diffusion model into an image-conditional shadow-free image generator. This approach enables robust, high-resolution shadow removal without an input shadow mask.
- We introduce a shadow-aware detail injection module that utilizes the VAE encoder features to modulate the pre-trained VAE decoders, selectively aligning per-pixel details from the input image with those in the output shadow-free image.
- Extensive experimental results on public datasets demonstrate that the proposed method significantly outperforms state-of-the-art shadow removal methods. Additionally, we show that our method exhibits better generalizability by training on one dataset and testing on another.

2. Related Work

2.1. Deep Shadow Removal

Deep learning methods have significantly advanced shadow removal, enabling end-to-end learning of a mapping from input images to corresponding shadow-free images, with or without the aid of shadow masks. For example, DeshadowNet [28] combines multi-level features to predict a shadow matte, thereby enhancing the efficiency of shadow removal. Hu et al. [1, 10] utilize direction-aware spatial context for precise shadow detection and elimination. Fu et al. [3] formulate the shadow removal task as a multiple exposure image fusion problem. BMNet [48] introduces a bijective mapping that integrates shadow removal with shadow generation, improving overall performance in shadow removal tasks.

More recently, ShadowFormer [4] presents a Swin-transformer-based shadow removal network that harnesses multi-scale attention to exploit contextual relationships between shadowed and non-shadowed regions. Meanwhile, DMTN [18] proposes a decoupled multi-task network that learns decomposed features specifically tailored for shadow removal. ShadowRefiner [2] introduces a mask-free model that integrates spatial and frequency domain representations for image shadow removal, achieving top results in the

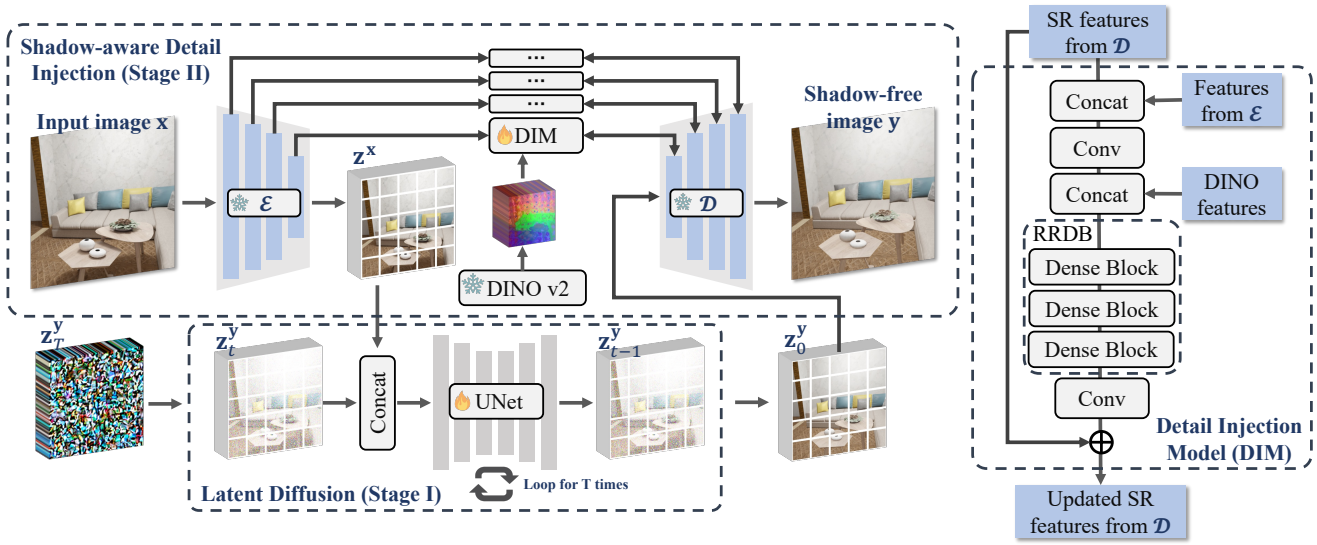

Figure 2. **Our proposed network.** We propose a two-stage shadow removal network based on Stable Diffusion (SD). (1) In the first stage, as shown in the bottom half, we fine-tune the pre-trained UNet in SD within the latent space defined by SD's pre-trained VAE ($\mathcal{E}$ and $\mathcal{D}$). We found that the pre-trained latent space can effectively represent shadow-free images. (2) In the second stage, as shown in the top half, we modulate the VAE decoder $\mathcal{D}$ by selectively adding features from the VAE encoder $\mathcal{E}$ using a Detail Injection Model (DIM). The model consists of multiple RRDB layers, which inject shadow-free texture details into the decoder features. With these two stages, our proposed network can generate high-quality, shadow-free images that preserve fine details.

Perceptual Track of the NTIRE 2024 Image Shadow Removal Challenge [36]. Xu et al. [41] propose a path-tracing pipeline to generate shadow removal data that considers indirect illumination in indoor scenes. They also introduce a semantics and geometry-aware network for efficient mask-free shadow removal. HomoFormer [40] designs random shuffle and inverse shuffle operations to create a homogenized appearance of shadows, allowing for effective processing by local self-attention layers.

2.2. Diffusion-based Shadow Removal

Diffusion models [9, 34] are generative models that learn data distributions through Gaussian noise blurring and reverse denoising. Initially developed for image generation, these models have become prominent in applications including image super-resolution [31, 38], depth/normal prediction [12, 42], and shadow removal [23]. In the context of shadow removal, ShadowDiffusion [5] introduces a diffusion model that generates both a shadow-free image and a refined shadow mask, using shadow degradation as a prior and employing an unrolling diffusion process to eliminate shadows. DeS3 [11] enhances ShadowDiffusion by incorporating an adaptive attention mechanism and ViT similarity loss, removing the dependence on shadow masks. Guo et al.[6] propose a diffusion-based unsupervised shadow removal pipeline that utilizes a boundary-aware divide and conquer strategy. Diff-Shadow[22] employs a globally guided diffusion model pipeline that focuses on the patch and global information, enabling shadow removal in images of arbitrary resolution. Liu et al. [20] first separate reflectance and illumination, then use a diffusion model to correct degraded lighting in shadowed areas, followed by the progressive restoration of texture details through an illumination-guided texture restoration module.

There are also methods that utilize a latent space to enhance diffusion-based shadow removal. Refusion [6] presents a U-Net-based latent diffusion model that utilizes mean-reverting stochastic differential equations for high-resolution image restoration tasks. The model performs image diffusion in a low-resolution latent space while preserving details through hidden-connections. However, being end-to-end trained on the shadow-removal dataset, it suffers from limited generalizability. Mei et al. [24] conduct diffusion-based shadow removal with a shadow mask at both the general and instance levels. They improve the diffusion process by conditioning on a learned latent feature space that captures the characteristics of shadow-free images and propose fusing noise features with the diffusion network to alleviate potential local optima.

Unlike existing diffusion-based methods, we introduce a two-stage fine-tuning strategy that fully utilizes the visual priors from the pre-trained Stable Diffusion (SD) model [29]. With our latent shadow removal and shadow-aware detail injection stages, our method demonstrates strong generalization capability while effectively preserv-

ing non-shadow details.

3. Proposed Method

3.1. Overview

Given a dataset of shadow and shadow-free image pairs $\langle \mathbf{x}, \mathbf{y} \rangle \in \mathcal{X}$, our goal is to train a network that can: 1. thoroughly eliminate shadows from the input image, and 2. preserve the content of all non-shadow areas. Although the first requirement focuses on the shadow (umbra) regions, the second is more concerned with non-shadow areas. These two objectives can be contradictory, making it challenging to design a single network that effectively addresses both. Therefore, we propose a two-stage shadow removal framework to robustly eliminate various types of shadows across different scenarios (Fig. 2).

In the **first stage**, as shown at the bottom of Fig. 2, we perform shadow removal in latent space (Sec. 3.2). We condition the Stable Diffusion (SD) model [29] using the latent code of the input shadow image $\mathbf{z}^{\mathbf{x}}$ and fine-tune a pre-trained SD model to generate the latent code of a shadow-free image $\mathbf{z}_0^{\mathbf{y}}$. As shown in Fig. 1, this latent code can be decoded into a coarse image that effectively eliminates shadows, although some details may be lost.

In the **second stage**, as illustrated at the top of Fig. 2, we perform detail injection using the encoder $\mathcal{E}$ and decoder $\mathcal{D}$ within SD (Sec.3.3). We fix the pretrained encoder and decoder and employ a set of Detail Injection (DI) models to integrate shadow-free details into the decoder features. These details are derived from each encoder layer. After training on pairs of shadow and shadow-free images, the DI models learn to inject details from the input image that are free of shadows.

3.2. Shadow Removal in Latent Space

Building on recent advancements in dense prediction tasks [7, 12, 42], we formulate the shadow removal task in the first stage as an image-conditioned annotation generation problem using SD. This approach conducts the diffusion process within a low-dimensional latent space generated by a pair of encoder and decoder $\langle \mathcal{E}, \mathcal{D} \rangle$, where $\mathcal{E}(\mathbf{x}) = \mathbf{z}^{\mathbf{x}}$ and $\mathcal{D}(\mathbf{z}^{\mathbf{x}}) \approx \mathbf{x}$.

Compared to the diffusion model [9], Stable Diffusion offers enhanced generation stability and training efficiency. However, given the limited number of shadow and shadow-free image pairs, we opt for a fine-tuning strategy rather than training a SD model from scratch. We directly utilize the pre-trained $\mathcal{E}$ and $\mathcal{D}$ in SD and focus on fine-tuning the denoiser U-Net, as we have found that the pre-trained $\mathcal{E}$ and $\mathcal{D}$ perform well for both shadow and non-shadow images, even though they were not specifically trained on shadow-free images (Fig. 3).

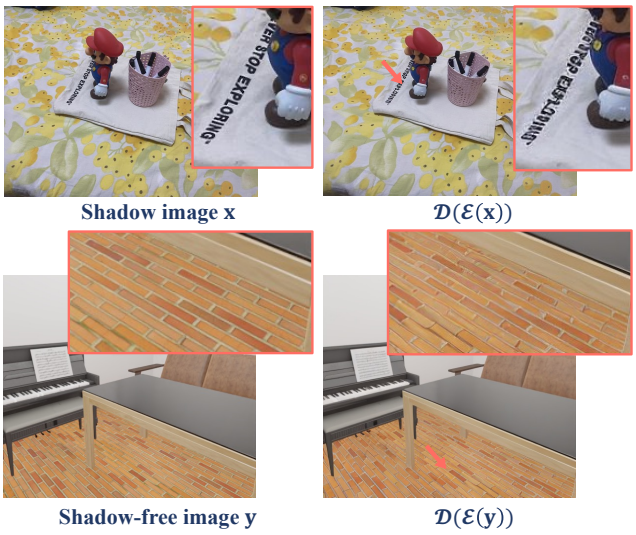

Figure 3. **Latent space in VAE.** We apply the pre-trained encoding and decoding process on a shadow image $\mathbf{x}$ or a shadow-free image $\mathbf{y}$. For a shadow-free image, this process does not introduce additional shadows, meaning the pre-trained latent space can also represent the shadow-free image. However, as shown in the zoomed-in view, some details, like text, may be lost in this process.

Fine-tuning Process. The fine-tuning for Stable Diffusion [9] relies on forward-noising and reverse-denoising processes that operate within the latent space. In the forward process, Gaussian noise is progressively added over time steps $t \in [1, T]$ to the sample $\mathbf{z}^{\mathbf{y}}$, generating a noisy sample $\mathbf{z}_{\mathbf{t}}^{\mathbf{y}}$:

$$\mathbf{z}_{\mathbf{t}}^{\mathbf{y}} = \sqrt{\overline{\alpha}_t}\mathbf{z}^{\mathbf{y}} + \sqrt{1 - \overline{\alpha}_t}\epsilon, \tag{1}$$

where $\epsilon \sim \mathcal{N}(0, I)$, $\overline{\alpha}_t := \prod_{s=1}^{t}(1 - \beta_s)$, and $\beta_1, \beta_2, \ldots, \beta_T$ defines the noise schedule over T steps. In the reverse process, a U-Net [30] denoted as f_θ takes the encoding features of the input image $\mathbf{z}^{\mathbf{x}}$ and the sample $\mathbf{z}_{\mathbf{t}}^{\mathbf{y}}$ as inputs, outputting a prediction of the clean sample $\hat{\mathbf{z}}^{\mathbf{y}}$:

$$\hat{\mathbf{z}}^{\mathbf{y}} = f_\theta(\mathbf{z}_{\mathbf{t}}^{\mathbf{y}}, \mathbf{z}^{\mathbf{x}}, t). \tag{2}$$

During fine-tuning, we randomly sampling a time step $t \in [1, T]$ and minimizing the corresponding loss function L_t:

$$\mathcal{L}_t = ||\mathbf{z}^{\mathbf{y}} - f_\theta(\mathbf{z}_{\mathbf{t}}^{\mathbf{y}}, \mathbf{z}^{\mathbf{x}}, t)||_2^2, \tag{3}$$

where $\mathbf{z}^{\mathbf{y}}$ is the encoding features of the ground-truth shadow-free image.

Denoising Process. During inference, we use DDIM [33], which implements an implicit probabilistic model that can significantly reduce the number of denoising steps while maintaining output quality. Formally, the denoising process

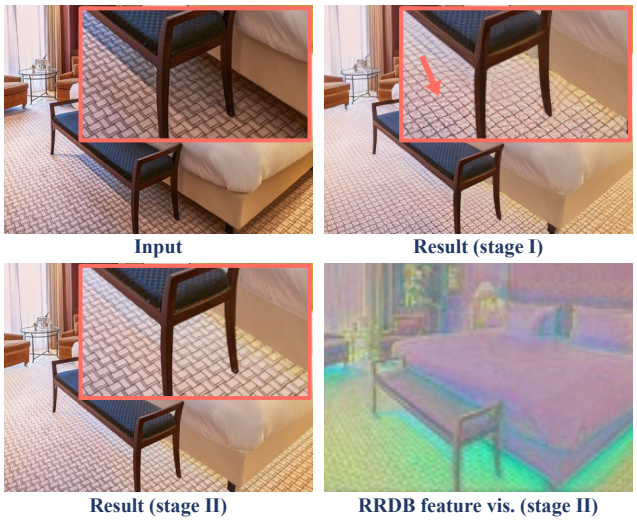

Figure 4. **Some details in the input image, such as the textures on the floor, change during latent space diffusion in stage one.** In stage two, our method preserves these high-quality details while effectively removing shadows. As shown in the bottom-left image, where we map the RRDB features to three dimensions using PCA, we observe that shadow areas have distinct features (in green), indicating that our detail injection model is shadow-aware.

from $\mathbf{z}_\tau^\mathbf{y}$ to $\mathbf{z}_{\tau-1}^\mathbf{y}$ is:

$$\mathbf{z}_{\tau-1}^\mathbf{y} = \sqrt{\bar{\alpha}_{\tau-1}}\hat{\mathbf{z}}_\tau^\mathbf{y} + \text{direction}(\mathbf{z}_\tau^\mathbf{y}) + \sigma_\tau \epsilon_\tau, \quad (4)$$

where $\hat{\mathbf{z}}_\tau^\mathbf{y}$ is the predicted clean sample at the denoising step τ, $\text{direction}(\mathbf{z}_\tau^\mathbf{y})$ represents the direction pointing to $\mathbf{z}_\tau^\mathbf{y}$. $\tau \in \{\tau_1, \tau_2, \ldots, \tau_S\}$ is an increasing sub-sequence of the time-step set $[1, T]$ used for fast sampling.

3.3. Shadow-aware Detail Injection

As illustrated in Fig. 4, while the output of the conditional Stable Diffusion is visually striking—especially in its ability to completely remove shadows—it often diverges from the ground truth, particularly in high-detail areas, such as the floor texture. We believe the primary reason for this is the loss of high-frequency information in the latent space. As shown in Fig. 3, directly decoding the encoded features $\mathcal{D}(\mathcal{E}(\mathbf{x}))$ also results in a loss of detail.

To eliminate shadows while preserving details, we introduce a shadow-aware detail injection step utilizing a set of DI models. Each model selectively extracts information from a specific intermediate layer of $\mathcal{E}$ to modulate the corresponding intermediate layers of $\mathcal{D}$. Let $\mathbf{e}_i$ and $\mathbf{d}_i$ denote the encoder and decoder features for the i-th intermediate layer. The DI model takes the concatenation of $\mathbf{e}_i$ and $\mathbf{d}_i$ as input, output a modulated feature $\tilde{\mathbf{d}}_i$:

$$\begin{aligned} \mathbf{d}_0 &= \mathcal{D}_0(\mathbf{z}_0^\mathbf{y}), \quad (5)\\ \tilde{\mathbf{d}}_i &= \mathbf{d}_i + \text{Conv}(\text{RRDB}(\text{Conv}(\mathbf{d}_i, \mathbf{e}_{n-i}(\mathbf{z}^\mathbf{x})), f_d(\mathbf{x}))),\\ \mathbf{d}_i &= \mathcal{D}_i(\tilde{\mathbf{d}}_{i-1}), i \in [1, 3], \hat{\mathbf{y}} = \text{Conv}(\tilde{\mathbf{d}}_3) \end{aligned}$$

Here, $\text{RRDB}(\cdot)$ refers to the Residual-in-Residual Dense Block module [46], which consists of three Dense Blocks. $\hat{\mathbf{y}}$ is prediected shadow-free image. $f_d(\mathbf{x})$ are features extracted using DINO v2 [26]. We concatenate the DINO features with $\mathbf{e}_i$ and $\mathbf{d}_i$ after passing through a convolutional layer to enhance the generalizability of the DI models.

Since $\mathbf{e}_i$ and $\mathbf{d}_i$ contain features from the input shadow image $\mathbf{x}$ and the shadow-free latent code $\mathbf{z}_0^\mathbf{y}$, respectively, our DI models can effectively identify shadow areas by utilizing both features as input. Although the latent codes may lose high-frequency details, they still provide valuable cues for shadow regions, which are primarily low-frequency. As illustrated in Fig. 4, where we visualize the features in the last layer of the RRDB module, we can observe that the shadow areas exhibit distinct feature responses. We present only the RRDB responses for $\mathbf{d}_2$. Please refer to the supplementary for additional results.

Losses. To train the DI models, we minimize a total loss $\mathcal{L}_{total}$, which combines an L1 loss and a perceptual loss measured using the Learned Perceptual Image Patch Similarity (LPIPS) metric [45] between the output image $\hat{\mathbf{y}}$ and the ground-truth shadow-free image $\mathbf{y}$:

$$\mathcal{L}_{total} = ||\hat{\mathbf{y}} - \mathbf{y}||_1 + ||\text{LPIPS}(\hat{\mathbf{y}}) - \text{LPIPS}(\mathbf{y})||_2^2. \quad (6)$$

3.4. Implementation Details.

Reducing variance. As shown in Eq. 3, in the LDM finetuning stage, we use $\mathbf{z}_0$-prediction rather than the standard ϵ-prediction parameterization for the denoising model. This choice, as recommended by Lotus[7], helps reduce variance during stochastic LDM inference. We compare these two approaches on the ISTD+ [14] dataset by running inference on the test set with five different random seeds (1, 2, 3, 4, 5). We calculate the PSNR variance across the five outputs for each image and obtain the average variance. As

Method	Mean↑	Variance↓
Our stage one (ϵ-pred)	29.66	0.239
Our stage one ($\mathbf{z}_0$-pred)	**29.95**	**0.146**
Our stage two	**35.02**	**0.160**
DeS3 [11]	31.33	1.075

Table 1. **Low variance.** In the first stage, our method achieves reduced variance by using $\mathbf{z}_0$-prediction. In the second stage, our method improves PSNR while maintaining low variance, resulting in robust, high-quality shadow removal results.

Method	Year	Mask-free	ISTD+ Dataset		SRD Dataset		INS Dataset		WSRD+ Dataset	
			PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑
AutoExposure [3]	2021	No	29.45	0.861	29.24	0.938	27.91	0.957	21.66	0.752
Zhu et al. [49]	2022	No	—	—	32.05	0.965	—	—	—	—
BMNet [48]	2022	No	33.98	0.972	31.97	0.965	27.90	0.958	24.75	0.816
ShadowFormer [4]	2023	No	35.46	0.971	32.90	0.958	28.62	0.963	25.44	0.820
DMTN [18]	2023	No	32.23	0.966	33.77	0.968	28.83	0.969	—	—
ShadowDiffusion [5]	2023	No	34.63	0.967	34.73	0.970	29.12	0.966	—	—
HomoFormer [40]	2024	No	35.72	0.977	34.36	0.977	28.98	0.965	—	—
InstanceShadow [24]	2024	No	34.69	0.968	31.61	0.959	28.16	0.948	—	—
TBRNet [19]	2023	Yes	31.91	0.964	31.83	0.953	—	—	—	—
Refusion [23]	2023	Yes	32.41	0.961	31.60	0.949	28.13	0.958	22.32	0.738
DeS3 [11]	2024	Yes	31.39	0.957	34.11	0.968	27.89	0.947	—	—
OmniSR [41]	2024	Yes	33.34	0.970	32.87	0.969	30.38	0.973	26.07	0.835
Ours	—	Yes	35.19	0.974	33.63	0.968	30.56	0.975	26.26	0.827

Table 2. **Quantitative comparisons on ISTD+, SRD, INS, and WSRD+ datasets.** Best results are highlighted as 1st, 2nd and 3rd.

shown in Table 1, using $\mathbf{z}_0$-prediction in our LDM fine-tuning stage yields higher PSNR and lower variance compared to ϵ-prediction. Additionally, we assess the average variance of our final results after the detail injection stage, and Table 1 shows that our final results have significantly lower variance than DeS3 [11], another mask-free shadow removal method based on diffusion models. It indicates that our generative shadow removal method is highly robust to randomness, which is a critical factor for practical use.

Training details. Our model is trained on a GPU server with four GeForce RTX 4090 GPUs using PyTorch 2.0.1 [27] with CUDA 11.7. We employ the Adam optimizer [13] for training. In the first stage, the LDM is fine-tuned for 40 epochs on INS [41] and 200 and 300 epochs on ISTD+ [14, 37] and SRD [28], respectively. We utilize a constant learning rate of 3×10^{-4} with a batch size 16. For the INS dataset, each batch consists of 512×512 patches, while for the other datasets, batches are composed of randomly cropped 480×480 patches.

In the second stage, we maintain the constant learning rate of 5×10^{-4} and batch size of 16. The batch shapes for each dataset are consistent with those in the first stage, and the number of epochs for training the detail injection models is 50 for INS and 300 and 200 for ISTD+ and SRD, respectively. For the DINO features [26], we resize the image to $\frac{14W}{16} \times \frac{14H}{16}$ using bilinear interpolation, extract the DINO features, and apply a 1×1 convolution layer to obtain a feature map of shape $256 \times \frac{W}{16} \times \frac{H}{16}$. To reduce memory footprint, we only concatenate the DINO features with the first two layers of the VAE decoder. Additional details can be found in the supplementary.

4. Experiments

Datasets. We conduct experiments on four shadow removal datasets: SRD [28], ISTD+ [14, 37], WSRD+ [35, 36], and INS [41]. The SRD dataset [28] includes 3,088 paired shadow and shadow-free images, divided into 2,680 for training and 408 for testing. The ISTD+ dataset [14, 37] provides 1,870 triplets of shadow images, shadow masks, and shadow-free images, with 1,330 used for training and 540 for testing. WSRD+[35, 36] contains 1,000 pairs of shadow and non-shadow images captured under both spotlight and diffuse illumination. Since the test data is not open-sourced, we use its 100 evaluation images for testing. Lastly, the INS dataset [41] comprises 30,000 synthetic training images under direct and indirect illumination, a synthetic test set of 2,000 images.

Evaluation Metrics. Following previous methods [3, 4, 15], we evaluated images with a resolution of 256×256. We report results using the Peak Signal-to-Noise Ratio (PSNR) and the Structure Similarity Index Measure (SSIM) [39], adopting the MATLAB evaluation codes as provided by Zhu et al. [49]. For the WSRD+ [35] dataset, we used its evaluation data and the evaluation code provided by the NTIRE 2024 Image Shadow Removal Challenge [36] for comparison.

4.1. Comparisons

We compare our method to twelve state-of-the-art shadow removal techniques, evaluating both quantitatively and qualitatively. Among these, eight methods require a shadow mask as input: AutoExposure [3], Zhu et al. [49], BMNet [48], ShadowFormer [4], DMTN [18], ShadowDiffusion [5], and HomoFormer [40]. The remaining four meth-

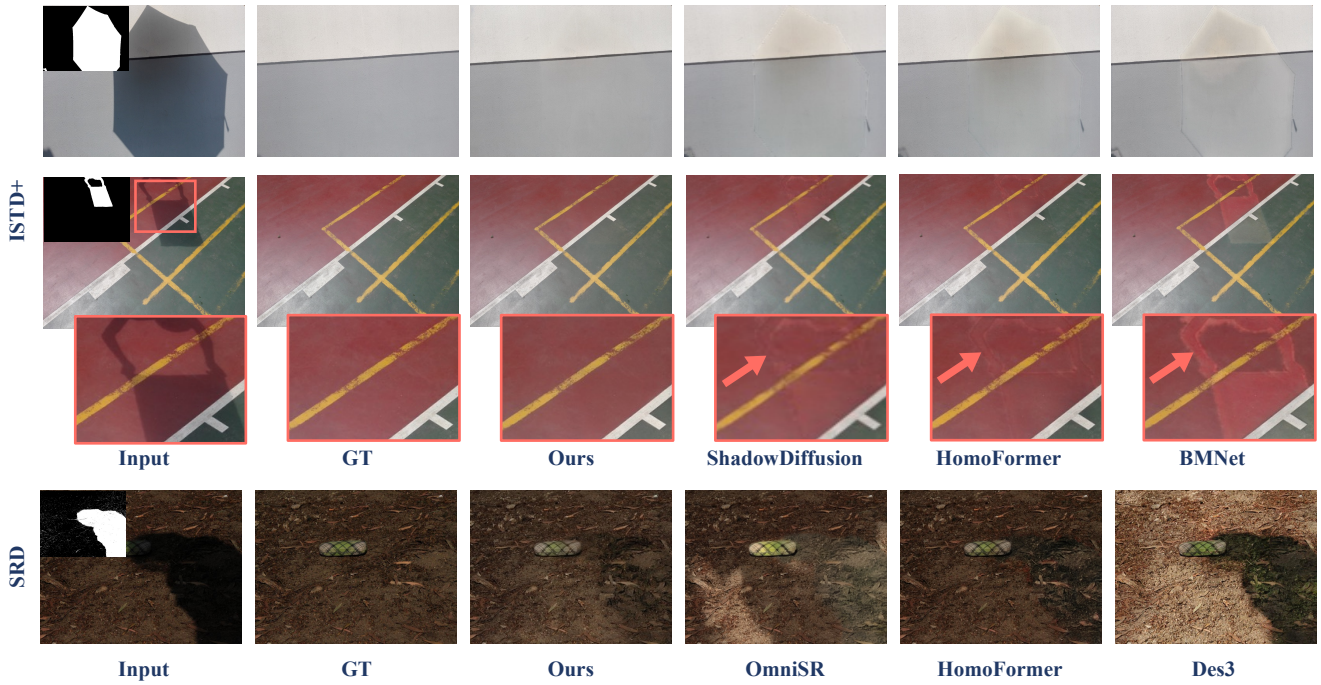

Figure 5. **Comparisons with SOTA shadow removal methods on the ISTD+ and SRD datasets.** The input mask, required by methods that are not mask-free, is shown in the top-left corner.

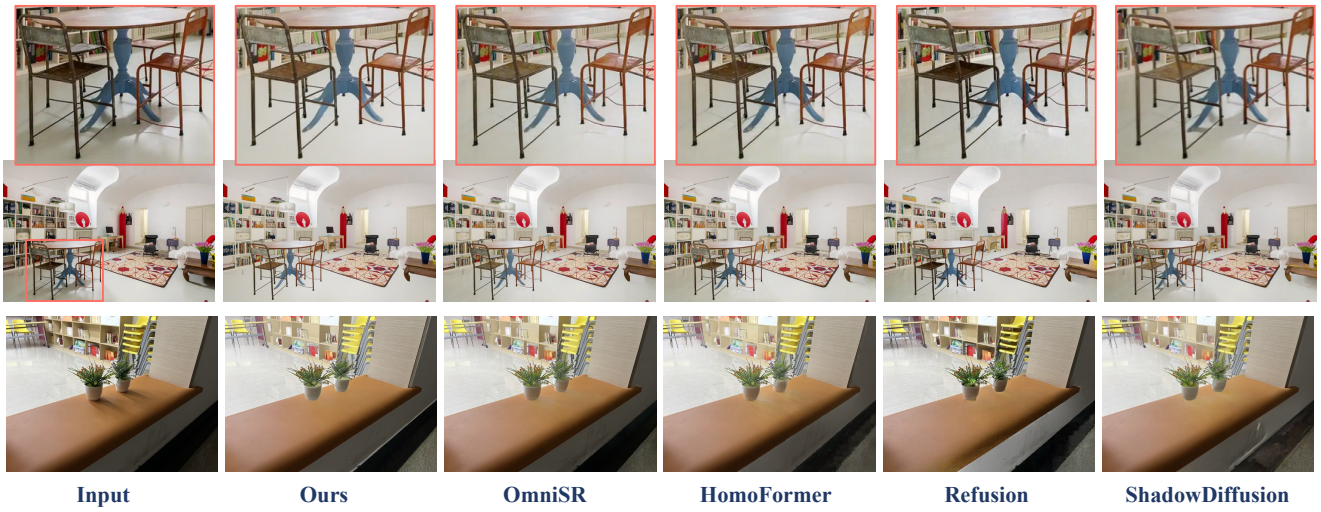

Figure 6. **Comparisons with SOTA shadow removal methods in real scenes.** We compare our method to OmniSR [41], HomoFormer [40], Refusion [23], and ShadowDiffusion [5], all trained on the INS dataset [41]. Results show that our approach achieves more comprehensive shadow removal, even in complex scenes.

ods—TBRNet [19], Refusion [23], Des3 [11], and OmniSR [41]—are mask-free, meaning they do not require a shadow mask as input. All comparisons use the original authors' implementations and hyperparameters, or the provided checkpoints or results, if available. For the INS dataset, we train these methods following the procedure in OmniSR [41].

For methods that are not mask-free, we use the shadow masks provided by the dataset as the default input. Since SRD [28] does not provide masks, we use the shadow masks generated by DHAN [1], which computes shadow matting from shadow/shadow-free pairs and generates binary

Method	ISTD+→SRD		SRD→ISTD+		INS→WSRD+	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
ShadowDiffusion	24.26	0.890	29.48	0.947	19.40	0.825
Refusion	22.56	0.886	27.31	0.924	18.73	0.787
DeS3	22.09	0.880	—	—	—	—
OmniSR	25.86	0.917	30.19	0.951	20.03	0.825
Ours	**26.23**	**0.923**	**30.23**	**0.953**	**20.11**	**0.828**

Table 3. **Cross-dataset evaluation.** ISTD+→SRD means training on the ISTD+ dataset and tesing on the SRD dataset.

Dataset	ISTD+	SRD	INS
	PSNR/SSIM	PSNR/SSIM	PSNR/SSIM
Full	**35.19/0.974**	**33.63/0.968**	**30.56/ 0.975**
W/o stage two	29.95/0.843	28.18/0.868	28.78/0.924
W/o DINO	34.76/0.971	33.04/0.965	30.30/0.975

Table 4. **Ablation studies.** W/o DINO: without DINO in the detail injection model.

masks through thresholding. In the case of the INS dataset, its shadow masks are represented as 0-1 probability maps rather than binary masks. Since it is a synthetic dataset with highly accurate masks, we found that using them as input results in methods producing outputs nearly identical to the ground truth (with PSNR values exceeding 50). To address this, following OmniSR [41], we use FDRNet [47] to detect shadow masks and input the detected masks instead.

As shown in Table 2, our method achieves the highest PSNR score on the INS dataset. On the ISTD+ dataset, our approach is the best mask-free method, performing only slightly below HomoFormer [40] and ShadowFormer [4], which use ground-truth shadow masks as input, and outperforming most other methods that also rely on ground-truth shadow masks. As shown in the first two columns of Fig. 5, although without mask as input, our method is capable of thoroughly removing shadows, even with limited semantic context. We think it is because of leveraging the generative priors in our latent diffusion model.

For the SRD dataset, our method also delivers competitive results compared to methods that utilize masks provided by DHAN [1]. Among shadow-free methods, our performance is second only to DeS3. As illustrated in Fig. 5, our method produces visually appealing results with nearly complete shadow removal. We found that our method achieves better shadow removal quality than DeS3; however, in areas with highly detailed textures, such as sand and grass, the PSNR is lower than that of DeS3.

On the INS dataset, our method achieves the highest PSNR and SSIM, as shown in Table 2. We also compare our method with others on real indoor scene images, all trained on the INS dataset. As illustrated in Fig. 6, our approach more effectively removes shadows from multiple light sources, such as those beneath the table and chair, as well as soft shadows behind the small planter and on the bookshelf in the background.

We further evaluated our method on high-resolution shadow removal using the WSRD+ [35] dataset, with images at 1920×1440 resolution. The first stage requires no modification. We resize each image to 640×480 for the LDM fine-tuning. In the second stage, we introduce a minor adjustment by adding one patch-partition layer and one patch-merging layer [21]. Additional details are provided in the supplementary materials. As shown in Table 2, our method outperforms others on the WSRD+ dataset, with more results available in the supplementary material. The relatively low PSNR scores for all methods on the WSRD+ dataset can be attributed to exposure differences between the input and ground-truth images.

4.2. Cross-dataset evaluation.

To evaluate the generalizability of our approach, we conduct a cross-dataset evaluation. Specifically, we train our model on the training set of one dataset and test it on the testing set of a different dataset. Typically, training and testing data within the same dataset share similar shadow patterns, backgrounds, or lighting conditions. This cross-dataset evaluation presents a more challenging test for shadow removal methods and provides a more stringent measure of generalizability. We use $A \rightarrow B$ to denote training on dataset A and testing on dataset B. As shown in Table 3, our method achieves the best results on ISTD+→SRD, SRD→ISTD, and INS→WSRD+. Here, we compare only with DeS3 [11] on ISTD+→SRD, as it provides checkpoints only for ISTD+ dataset. Its open-source code is a basic version that excludes certain components, including the VIT loss. For the INS→WSRD+ experiment, high-resolution images from the WSRD+ dataset are resized to 640×480 for inference and evaluation. Qualitative comparisons for this evaluation are provided in the supplementary material.

4.3. Ablation study.

To validate our model design, we conducted ablation studies on the two-stage pipeline and the addition of DINO features. The results are reported in Table 4. Stage two plays a critical role in our method, as removing it leads to a significant drop in PSNR and SSIM. This is due to stage one's inability to effectively preserve high-frequency details. Additionally, incorporating DINO features into the detail injection model of stage two proves beneficial, enhancing PSNR and SSIM across all three datasets. For qualitative results related to the ablation study, please refer to the supplementary materials.

5. Conclusion

This paper introduces a novel detail-preserving latent diffusion pipeline that effectively eliminates complex shadows while preserving shadow-free details. The proposed method leverages the rich visual priors of a pre-trained Stable Diffusion model and propose a two-stage fine-tuning strategy to adapt the SD model for stable and efficient shadow removal. Extensive experiments demonstrate that our approach outperforms state-of-the-art shadow removal methods, particularly in terms of generalizability.

Limitation and future work. Our method may still miss some shadows in complex scenes, such as the self-shadows on the flower pot in Fig. 6. In the future, we plan to incorporate unsupervised or self-supervised signals to further enhance the generalizability of our method.

References

[1] 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*, pages 10680–10687, 2020. 2, 7, 8

[2] Wei Dong, Han Zhou, Yuqiong Tian, Jingke Sun, Xiaohong Liu, Guangtao Zhai, and Jun Chen. Shadowrefiner: Towards mask-free shadow removal via fast fourier transformer. *arXiv preprint arXiv:2406.02559*, 2024. 2

[3] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10571–10580, 2021. 2, 6

[4] Lanqing Guo, Siyu Huang, Ding Liu, Hao Cheng, and Bihan Wen. Shadowformer: Global context helps image shadow removal. *arXiv preprint arXiv:2302.01650*, 2023. 1, 2, 6, 8

[5] Lanqing Guo, Chong Wang, Wenhan Yang, Siyu Huang, Yufei Wang, Hanspeter Pfister, and Bihan Wen. Shadowdiffusion: When degradation prior meets diffusion model for shadow removal. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14049–14058, 2023. 1, 3, 6, 7

[6] Lanqing Guo, Chong Wang, Wenhan Yang, Yufei Wang, and Bihan 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

[7] Jing He, Haodong Li, Wei Yin, Yixun Liang, Leheng Li, Kaiqiang Zhou, Hongbo Liu, Bingbing Liu, and Ying-Cong Chen. Lotus: Diffusion-based visual foundation model for high-quality dense prediction. *arXiv preprint arXiv:2409.18124*, 2024. 4, 5

[8] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 1

[9] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 3, 4

[10] 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/CVF international conference on computer vision*, pages 2472–2481, 2019. 2

[11] Yeying Jin, Wei Ye, Wenhan Yang, Yuan Yuan, and Robby T Tan. Des3: Adaptive attention-driven self and soft shadow removal using vit similarity. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2634–2642, 2024. 3, 5, 6, 7, 8

[12] Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9492–9502, 2024. 1, 2, 3, 4

[13] D. Kingma and J. Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2015. 6

[14] 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. 1, 5, 6

[15] Hieu Le and Dimitris 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. 6

[16] Zhengqi Li and Noah 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

[17] Zhengqin Li, Jia Shi, Sai Bi, Rui Zhu, Kalyan Sunkavalli, Miloš Hašan, Zexiang Xu, Ravi Ramamoorthi, and Manmohan Chandraker. Physically-based editing of indoor scene lighting from a single image. In *European Conference on Computer Vision*, pages 555–572. Springer, 2022. 1

[18] Jiawei Liu, Qiang Wang, Huijie Fan, Wentao Li, Liangqiong Qu, and Yandong Tang. A decoupled multi-task network for shadow removal. *IEEE Transactions on Multimedia*, 2023. 2, 6

[19] Jiawei Liu, Qiang Wang, Huijie Fan, Jiandong Tian, and Yandong Tang. A shadow imaging bilinear model and three-branch residual network for shadow removal. *IEEE Transactions on Neural Networks and Learning Systems*, 2023. 6, 7

[20] Yuhao Liu, Zhanghan Ke, Ke Xu, Fang Liu, Zhenwei Wang, and Rynson WH Lau. Recasting regional lighting for shadow removal. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3810–3818, 2024. 3

[21] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 2, 8

[22] Jinting Luo, Ru Li, Chengzhi Jiang, Mingyan Han, Xiaoming Zhang, Ting Jiang, Haoqiang Fan, and Shuaicheng Liu.

Diff-shadow: Global-guided diffusion model for shadow removal. *arXiv preprint arXiv:2407.16214*, 2024. 2, 3

[23] Ziwei Luo, Fredrik K Gustafsson, Zheng Zhao, Jens Sjölund, and Thomas B Schön. Refusion: Enabling large-size realistic image restoration with latent-space diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1680–1691, 2023. 2, 3, 6, 7

[24] Kangfu Mei, Luis Figueroa, Zhe Lin, Zhihong Ding, Scott Cohen, and Vishal 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. 1, 3, 6

[25] Thomas Nestmeyer, Jean-François Lalonde, Iain Matthews, and Andreas Lehrmann. Learning physics-guided face relighting under directional light. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5124–5133, 2020. 1

[26] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 5, 6

[27] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017. 6

[28] 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. 1, 2, 6, 7

[29] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 1, 3, 4

[30] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015. 4

[31] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *IEEE transactions on pattern analysis and machine intelligence*, 45(4):4713–4726, 2022. 3

[32] Andres Sanin, Conrad Sanderson, and Brian C Lovell. Improved shadow removal for robust person tracking in surveillance scenarios. In *2010 20th International Conference on Pattern Recognition*, pages 141–144. IEEE, 2010. 1

[33] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 4

[34] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021. 3

[35] Florin-Alexandru Vasluianu, Tim Seizinger, and Radu Timofte. Wsrd: A novel benchmark for high resolution image shadow removal. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1826–1835, 2023. 1, 2, 6, 8

[36] Florin-Alexandru Vasluianu, Tim Seizinger, Zhuyun Zhou, Zongwei Wu, Cailian Chen, Radu Timofte, Wei Dong, Han Zhou, Yuqiong Tian, Jun Chen, et al. Ntire 2024 image shadow removal challenge report. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6547–6570, 2024. 3, 6

[37] 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. 6

[38] Jianyi Wang, Zongsheng Yue, Shangchen Zhou, Kelvin C.K. Chan, and Chen Change Loy. Exploiting diffusion prior for real-world image super-resolution. 2024. 3

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

[40] Jie Xiao, Xueyang Fu, Yurui Zhu, Dong Li, Jie Huang, Kai Zhu, and Zheng-Jun Zha. Homoformer: Homogenized transformer for image shadow removal. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 25617–25626, 2024. 1, 3, 6, 7, 8

[41] Jiamin Xu, Zelong Li, Yuxin Zheng, Chenyu Huang, Renshu Gu, Weiwei Xu, and Gang Xu. Omnisr: Shadow removal under direct and indirect lighting. *arXiv preprint arXiv:2410.01719*, 2024. 3, 6, 7, 8

[42] Chongjie Ye, Lingteng Qiu, Xiaodong Gu, Qi Zuo, Yushuang Wu, Zilong Dong, Liefeng Bo, Yuliang Xiu, and Xiaoguang Han. Stablenormal: Reducing diffusion variance for stable and sharp normal. *arXiv preprint arXiv:2406.16864*, 2024. 1, 2, 3, 4

[43] Weicai Ye, Shuo Chen, Chong Bao, Hujun Bao, Marc Pollefeys, Zhaopeng Cui, and Guofeng Zhang. Intrinsicnerf: Learning intrinsic neural radiance fields for editable novel view synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 339–351, 2023. 1

[44] Ruo Zhang, Ping-Sing Tsai, James Edwin Cryer, and Mubarak Shah. Shape-from-shading: a survey. *IEEE transactions on pattern analysis and machine intelligence*, 21(8): 690–706, 1999. 1

[45] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 5

[46] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual dense network for image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2472–2481, 2018. 5

[47] Lei Zhu, Ke Xu, Zhanghan Ke, and Rynson WH Lau. Mitigating intensity bias in shadow detection via feature decomposition and reweighting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4702–4711, 2021. 8

[48] Yurui Zhu, Jie Huang, Xueyang Fu, Feng Zhao, Qibin Sun, and Zheng-Jun Zha. Bijective mapping network for shadow removal. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5627–5636, 2022. 2, 6

[49] Yurui Zhu, Zeyu Xiao, Yanchi Fang, Xueyang Fu, Zhiwei Xiong, and Zheng-Jun Zha. Efficient model-driven network for shadow removal. In *Proceedings of the AAAI conference on artificial intelligence*, pages 3635–3643, 2022. 6