Shadow Removal using Denoising Diffusion Probabilistic Models

Abhinav Sharma
Stony Brook University

Sugam Dipak Devare
Stony Brook University

1. Abstract

Shadow removal aims to recover the image under the shadow region. Previous methods focused on solving a physical model of image composition. Diffusion models have recently emerged as powerful tools for generative modeling and can be applied to a variety of tasks, including image denoising, inpainting, super-resolution and image generation. The essential idea of a diffusion model is to systematically and slowly destroy structure in a data distribution through an iterative forward diffusion process. We then learn a reverse diffusion process that restores structure in data, yielding a highly flexible and tractable generative model of the data. We propose to explore the possibility of applying the diffusion method for shadow removal tasks. Since, shadow removal requires recovering the content in the shadow region rather than regenerating the region under the shadow, we explore ways to improve the efficiency of the algorithm.

2. Introduction

Shadows are created when an object blocks a light source, and they can make it difficult for computer vision algorithms to accurately segment, track, or recognize objects in an image. Shadows can be especially problematic because their edges can be similar in appearance to edges caused by changes in the material of an object. Additionally, dark regions in an image that are caused by a low albedo (i.e., the ability of a surface to reflect light) can be mistaken for shadows. As a result, many techniques have been developed to identify and remove shadows from images.

There are some methods that use pre-existing shadow detection techniques[1] or rely on user input[2] to obtain binary shadow masks, which can then be used to remove shadows from images. However, these methods can be problematic because it can be difficult to accurately detect light or soft shadows, and because supervised methods for shadow removal often lack generalization beyond the relatively small training datasets that they are based on.

Existing unsupervised methods (e.g. [3], [4], [5]) for identifying and removing shadows from images are typically based on generative adversarial networks (GANs) [6], and they require a large number of unpaired shadow and shadow-free images for training. These methods often rely on the statistical similarity between the shadow and shadow-free domains, but they can produce unrealistic or fake results when the two domains are statistically different. Additionally, training GANs can be unstable.

In this project, we propose a solution for shadow removal using Denoising Diffusion Probabilistic Models (DDPMs) [7][8]. We focus on two main methods. The first method involves using a binary mask to inpaint the shadow region in the image, effectively removing the shadow. The second method involves using intermediate-conditional sampling from the latent space of the DDPM to remove shadows from the image. By using these two approaches, we aim to develop a more effective and generalizable method for shadow removal.

3. Proposed Method

(a) Shadow Image (b) Shadow-Free Image

Figure 1. Shadow Removal

3.1. Overview

We divided the solutions into two board categories -

1. Shadow Removal using Shadow Mask - We use the shadow mask and model the shadow removal task as a conditional inpainting task.
2. Shadow Removal without using Mask - Here, we use the novel intermediate-conditional sampling method for shadow removal.

3.2. Dataset

The ISTD dataset contains 1870 triplets of shadow images, shadow-free images, and shadow masks. The dataset is divided into 1330 triplets for training and 540 triplets for testing.

3.3. Diffusion Models

Denoising Diffusion Probabilistic Model (DDPM) is trained on a nonshadow image $\mathbf{x}_0 \sim q(\mathbf{x}_0)$ via a forward diffusion process $q(\mathbf{x}_t \mid \mathbf{x}_{t-1})$ that sequentially adds the Gaussian noise at every time steps t, i.e., $q(\mathbf{x}_t \mid \mathbf{x}_{t-1}) = \mathcal{N}\left(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t \mathbf{I}\right)$, where $\{\beta\}_{t=0}^{T}$ is a variance schedule. The forward diffusion process follows a Markov chain of length T expressed as:

$$q(\mathbf{x}_{1:T} \mid \mathbf{x}_0) = \prod_{t=1}^{T} q(\mathbf{x}_t \mid \mathbf{x}_{t-1})$$

DDPM learns to reverse the process with the Gaussian transitions, i.e., $p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \boldsymbol{\mu}_\theta(\mathbf{x}_t, t), \boldsymbol{\Sigma}_\theta(\mathbf{x}_t, t))$ at each time step. The reverse denoising (generative) process is parameterized by a trainable network (e.g., U-Net), which estimates the mean $\boldsymbol{\mu}_\theta(\mathbf{x}_t, t)$ and the variance

$\mathbf{\Sigma}_\theta(\mathbf{x}_t, t)$ with parameter θ. This reverse process starts from a standard normal distribution $p(\mathbf{x}_T) = \mathcal{N}(\mathbf{x}_T; \mathbf{0}, \mathbf{I})$ and follows

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

The DDPM model is optimized by considering the variational lower bound. The noise prediction network can be optimized by this objective function

$$L_\theta(\mathrm{DM}) = \mathbb{E}_{\mathbf{x}_0, \boldsymbol{\epsilon}, t}\left[\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)\|^2\right]$$

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

$$\mathbf{x}_{t-1} = \boldsymbol{\mu}_\theta(\mathbf{x}_t, t) + \mathbf{\Sigma}_\theta^{1/2}(\mathbf{x}_t, t)\,\boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$$

4. Experiments

4.1. Exp 1 : RePaint with Places2 model

RePaint [9] is a denoising diffusion probabilistic model (DDPM) based inpainting approach that is effective even for extreme masks (Figure 2). It uses a pre-trained unconditional DDPM as the generative prior and alters the reverse diffusion iterations to condition the generation process using the given image information. This allows the model to produce high-quality and diverse output images for any inpainting task without modifying or conditioning the original DDPM network. First, we inverted the masks present in the ISTD dataset since RePaint required the region to be replaced to be black and the rest of the image mask white. Before training a diffusion model on the shadow dataset, we wanted to check if the shadow-removal task could be achieved using a pretrained model. We used a model training on the Place2 dataset, which has over 10 million images comprising 400+ unique scene categories.

The intuition behind this was that the shadow removal task is usually applied to images of general places, and using a model trained on a diverse dataset would give good results. It is evident from the results that the model hallucinates the region under the shadow based on properties from the rest of the image producing unrealistic or fake results. In some cases, the model did remove the shadow but could not recover the underlying textures and objects. The model worked well in places with flat and uniform textures.

4.2. Exp 2 : RePaint with Improved Diffusion Model

We use Denoising Diffusion Probabilistic Models as our generative model. We used the improved diffusion codebase to train the model on non shadow images with a goal to regenerate the original image. We training the model using the following hyperparameters -

- num_channels = 128
- num_res_blocks = 2
- learn_sigma = True
- diffusion_steps = 1000
- noise_schedule = linear
- attention_resolutions = 16, 8

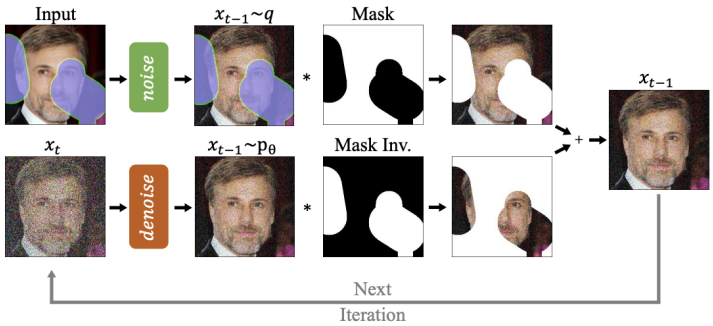

Figure 2. RePaint Architecture

In order to use the RePaint model, which is based on an older version of the guided diffusion architecture, we had to update the U-Net architecture to the latest version of the improved diffusion codebase. This involved making many changes to the code and configuration. The model then denoises the image by starting from pure noise and gradually filling in the unknown parts using the known parts of the image. Since diffusion models have semantic information encoded in their latent representation, this model works well at regenerating the region under the mask. The results are better than those from the pretrained Places2 model. When we used the shadow masks from the ISTD dataset, the outputs had a clear shadow boundary, so we updated the mask by extending the shadow by 3 pixels in all directions to remove this boundary.

4.3. Exp 3 : Intermediate-Conditional Sampling

The above two approaches require a mask of the shadow region. This requires manual cropping or shadow detection models, which might sometimes fail giving false positive outputs. We introduce a novel method for shadow removal called Intermediate-Conditional sampling. We use the same model trained on shadow-free images mentioned in the previous experiment.

During the sampling process, rather than starting from pure noise, start with a shadow image with some amount of noise added. The noise to be added is determined by the timestamp t from which we start the denoising process. For example, if the model is trained with 1000 diffusion steps, during sampling, we do not start from the noise at the 1000th step rather, we start from a noisy version of the shadow image at step 400. The model works very well in images with soft shadows (Figure 3).

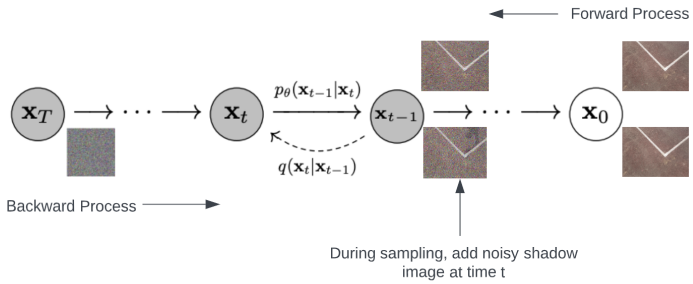

Figure 3. Intermediate-Conditional Sampling Method

5. Results

5.1. RePaint with Places2 model

Figure 4. Results from RePaint with Places2 model

Based on the results shown in Figure 4, we see that the RePaint system with the pre-trained Places2 model performs relatively well on inputs with good resolution or low shadow area. This is likely due to the fact that these images contain more semantic information, which the inpainting model can use to fill in the shadowed areas. However, the model performs poorly on inputs with complex patterns under shadow, low resolution, or large shadow areas.

5.2. RePaint with Improved Diffusion Model

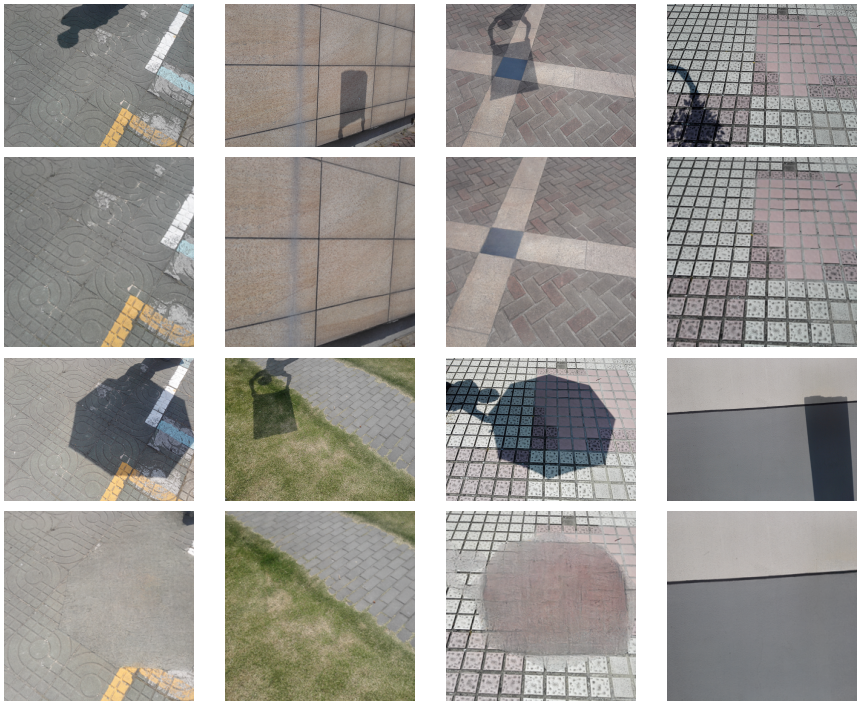

Figure 5. Results from RePaint with Improved Diffusion Model

The results presented in Figure 5 demonstrate the effectiveness of using RePaint in conjunction with the diffusion model trained on shadow-free images. This approach produces output images that are of higher quality and contain more semantic information compared to those produced by the Places2 pretrained model. While the model performs well on soft shadows, it struggles to accurately reconstruct more complex patterns and textures in the presence of hard shadows. Despite this limitation, the results obtained using the RePaint and diffusion model combination are very close to the ground-truth shadow-free images and represent a significant improvement over the Places2 pretrained model. Overall, these results suggest that the use of RePaint and a diffusion model trained on shadow-free images is a promising approach for removing shadows and improving image quality in a variety of applications.

5.3. Intermediate-Conditional Sampling Method Output

Figure 6. Results from Intermediate-Conditional Sampling

The results presented in Figure 6 demonstrate that starting the denoising process with samples that have added Gaussian noise (based on sampling timestamp) produces output images that are of higher quality and more semantically similar to the ground-truth than those produced with pure noise input images. This is likely because the intermediate sampling approach helps to retain some of the semantic information present in the shadowed image, which can be used to guide the denoising process and improve the output quality. However, the model still struggles to produce satisfactory results on inputs with large shadowed areas that cover most of the image. The output images generated by this approach are often softer and lack some of the fine details present in the ground-truth images and hence performs worse when compared to RePaint with Improved Diffusion Model on different evaluation metrics.

5.4. Metrics

Model	RMSE ↓	SSIM ↑	PSNR ↑
RePaint with Shadow Diffusion	9.56	0.923	30.57
Intermediate-Conditional Sampling	14.54	0.725	27.72

Table 1. Comparison between all the models

To evaluate the performance of our model, we use three metrics: root mean square error (RMSE), peak signal-to-noise ratio (PSNR), and structural similarity (SSIM). RMSE measures the difference between the output image produced by the model and the ground-truth shadow-free image. A lower RMSE indicates a better output. PSNR is a measure of the quality of a reconstructed signal, calculated by comparing the mean squared error of the reconstructed signal to the maximum possible power of the original signal. A higher PSNR means a higher quality of the reconstructed signal. SSIM compares the luminance, contrast, and structure of the images being compared, producing a score between -1 and 1, with a higher score indicating higher structural similarity. Table 1 shows the metrics for output generated using Intermediate-Conditional Sampling. Only outputs that are semantically similar to ground-truth shadow-free images have been included for metric calculation, i.e., around 200 test images.

5.5. Code

GitHub - https://github.com/sugamxp/shadow-diffusion

6. Future Scope

Future work can include adding structural features that help recover semantically meaningful structures during reverse diffusion. Moreover, categorizing the shadows into soft and hard shadows and adding a classification loss during training. Additionally, we can try to maintain the consistency of colours in the regions where the shadows have been removed. Finally, training the model on shadow images and conditioning it to predict the start image (shadow-free) instead of the noise at each reverse diffusion step. (Note - we tried this but faced issues during sampling - the output was noisy).

7. Conclusion

In this project, we have reproduced the results for RePaint, an in-painting system based on the diffusion model, RePaint using Places2 pre-trained model, cannot work on inputs that have texture or pattern under the shadow because the model hallucinates the region under the shadow based on properties from rest of the image producing unrealistic or fake results. We have updated RePaint's model architecture to work with the improved diffusion model trained on shadow-free images. This model works mains the underlying structure in the case of soft shadows. The results are better than those from the pre-trained RePaint model. We propose a novel model using intermediate-conditional sampling that does not rely on shadow masks and works well on soft shadows retaining underlying textures and patterns. The results of this model are soft due to lower reverse diffusion steps.

References

[1] Lei Zhu, Zijun Deng, Xiaowei Hu, Chi-Wing Fu, Xuemiao Xu, Jing Qin, and Pheng-Ann Heng. Bidirectional feature pyramid network with recurrent attention residual modules for shadow detection. In Proceedings of the European Conference on Computer Vision (ECCV), pages 121–136, 2018.

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

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

[4] Yeying Jin, Aashish Sharma, and Robby T Tan. Dcshadownet: 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.

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

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

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

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

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