LightIt: Illumination Modeling and Control for Diffusion Models

Peter Kocsis[1*]   Julien Philip[2]   Kalyan Sunkavalli[2]   Matthias Nießner[1]   Yannick Hold-Geoffroy[2]

[1]Technical University of Munich   [2]Adobe Research

peter-kocsis.github.io/LightIt/

Figure 1. **Teaser.** We present *LightIt*, a method for explicit lighting control of text-guided image generation. Given a normal map of the desired geometry and light direction with a solid angle, we introduce a method to predict direct shading, which is then used to generate high-quality images with coherent lighting.

Abstract

We introduce LightIt, a method for explicit illumination control for image generation. Recent generative methods lack lighting control, which is crucial to numerous artistic aspects of image generation such as setting the overall mood or cinematic appearance. To overcome these limitations, we propose to condition the generation on shading and normal maps. We model the lighting with single bounce shading, which includes cast shadows. We first train a shading estimation module to generate a dataset of real-world images and shading pairs. Then, we train a control network using the estimated shading and normals as input. Our method demonstrates high-quality image generation and lighting control in numerous scenes. Additionally, we use our generated dataset to train an identity-preserving re-lighting model, conditioned on an image and a target shading. Our method is the first that enables the generation of images with controllable, consistent lighting and performs on par with specialized relighting state-of-the-art methods.

1. Introduction

Generative imaging has significantly evolved over recent years. Diffusion models have shown outstanding capabilities to learn strong priors on large-scale real-image datasets. When used in conjunction with joint language-image embeddings, they have been successfully used for high-quality text-driven image generation. However, as these methods do not model light transport explicitly, the lighting in the produced images is uncontrollable and often inconsistent [3, 9]. However, lighting is an essential component of most images; without explicit control, artists rely on tedious and uncertain prompt engineering to try to control it.

Recent methods [13, 24, 50] have introduced control for various aspects of the generated images, for instance, depth or normals can be used to guide the geometry. Fine-grained control over the placement of generated objects can already be achieved [23, 37] and diffusion inversion enables modification of generated images [38, 43]. However, none of the methods can provide consistent and explicitly controllable lighting, which is the essence of photo-realistic images.

Diffusion models achieve incredible performance when trained on large datasets. However, the lack of real-world

lighting datasets is a major obstacle hindering progress on lighting control. Obtaining the lighting of a scene is a time-consuming task, requiring the decomposition of its appearance into lighting and material properties. However, we hypothesize that diffusion models do not require fine-grained lighting information, thus simplifying the required decomposition. Our analysis demonstrates that estimated single-bounce shading maps provide sufficient information and can be automatically obtained from real-world images.

In this work, we propose a single-view shading estimation method to generate a paired image and shading dataset. Given a single input image—either captured or generated—our model predicts a 3D density field, in which we trace rays toward a light to obtain cast shadows. Together with the estimated normals, giving us the cosine term, we predict single-bounce shading maps. This method notably allows us to generate shading maps for arbitrary lighting directions from a single image. Given outdoor panoramas, from which we can obtain the light direction, we thus generate a paired dataset of images and their shading. This dataset enables us to provide lighting control to the image generation process, which we also condition on normals to guide geometry.

As an additional application of our proposed illumination model, we further propose a relighting module conditioned on an input image and a target shading. Thanks to the strong natural image prior of Stable Diffusion (SD) [34], we obtain better generalization to real-world samples compared to methods from the literature trained on synthetic data. In summary, our main contributions are:

- We generate a paired image-shading dataset using our single-view, density field-based lighting estimation model, enabling single-bounce shading estimation for arbitrary lighting directions.
- We introduce lighting conditioning for controllable and coherent image generation using a diffusion-based model.
- We propose an identity-preserving relighting diffusion module utilizing the image prior for better generalization.

2. Related Work

Lighting controlled image generation. Generative imaging is a recent field that started to receive attention with the invention of GANs [7, 31]. However, one of the main challenges is the lack of control over the image generation process. To address this issue, methods such as StyleGAN [16] have been developed to provide control handles.

Recently, diffusion-based models were proposed to perform generative imaging [11, 29, 34, 38], enabling photoreal outputs from text prompts and democratizing generative imaging to the masses. As artists experimented with this new tool and explored its capabilities [49], the need for more control over the generation process arose. In particular, ControlNet [50] and T2I-Adapter [24] were recently proposed to allow users to control the generated image us-

ing a variety of modalities at the cost of image quality [18]. However, no approach exists for explicitly controlling the lighting of the generated imagery.

Relighting. Image relighting has traditionally been performed using classical approaches such as image-based rendering [4] or shape from shading [1]. The emergence of deep learning brought novel relighting approaches, initially using style transfer [6, 15, 22] or image-to-image translation [14, 39, 51]. Specialized relighting methods have begun with Xu et al. [47] that learns a relighting function from five images captured under predetermined illumination. Sengupta et al. [36] proposes to replace the traditional acquisition techniques with a regular monitor and camera setup. We encourage the reader to read the excellent review in [40] about rendering-based relighting.

Closer to our work, scene relighting methods both multi-view [26–28] and single-view [8, 48] generally use a combination of geometric and shading priors with a neural network to produce relit results. Outdoor NeRF-based relighting methods [35, 45] have been recently proposed, bringing the power of this implicit volumetric representation to relighting. Close in spirit to our shading model, OutCast [8] proposes to use depth and a large 3D CNN to process depth features sampled in image space to implicitly predict ray intersection. Our method builds on several of these ideas to propose a scene relighting approach combining volumetric scene representation, and explicit shadow ray-marching with diffusion-based image generation.

3. Method

Our method adds lighting control to the image generation process of a diffusion-based model. We develop a shading estimation method (Sec. 3.1) and generate a dataset of paired real images and shading maps (Sec. 3.4) to train a control module for SD [34] (Sec. 3.2). Our dataset enables additional applications, such as relighting (Sec. 3.3).

3.1. Shading estimation

To control the illumination of generated images, we aim to provide lighting information to the diffusion model. Estimating the shading from a single image is a challenging task even in the presence of depth estimation [8]. Inspired by Outcast [8], we develop a lightweight model to estimate *direct shading*, i.e. single-bounce illumination, from a single input image, which provides information about both shading and cast shadows. We show this pipeline in Fig. 2.

Specifically, we train a shading estimation model, which takes an image, a light direction, and a solid angle as input. A small 2D CNN (FeatureNet) first encodes the image to obtain a set of features. Then, using a pre-trained depth estimator (Sec. 3.4), we unproject these features to a multi-plane representation in Normalized Device Coordinates (NDC). Given a pixel's depth, the features are linearly
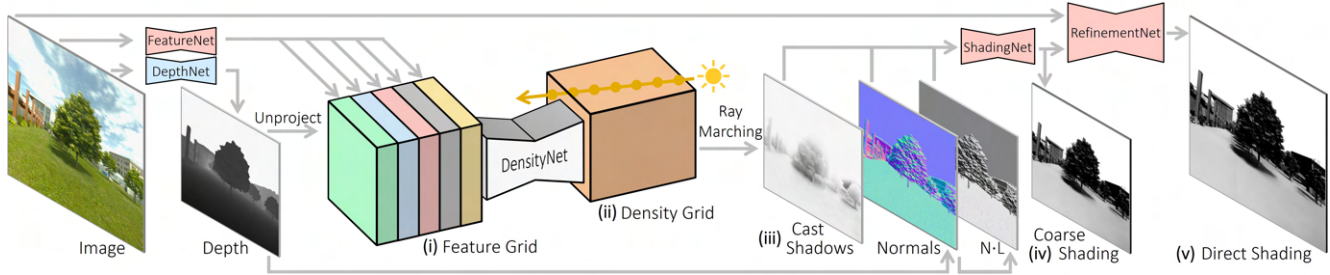
Figure 2. **Shading Estimation**. We estimate the direct shading of a single image. (i) We predict image features (FeatureNet) and unproject them to a 3D feature grid in NDC space. (ii) We predict a density field from the features (DensityNet). (iii) Given the sun's direction and solid angle, we trace rays toward the lightsource to obtain a coarse shadow map. (iv) Using the shadows and N-dot-L shading information, we predict a coarse shading map (ShadingNet). (v) We refine the shading map to get our direct shading (RefinementNet).

distributed between the two planes closest to the depth.

A small 3D CNN (DensityNet) processes the multi-plane and predicts a 3D density field. We render a cast shadow map for the light direction and angle using volumetric ray-marching. Finally, a 2D CNN (ShadingNet) transforms this shadow map and an N-dot-L cosine term map into coarse direct shading. To further improve the shading estimation quality, we apply a refinement module, which uses the input image and the predicted coarse shading.

We train our model on synthetic pairs of rendered images and shadings using an l_2 loss. To better guide the training, we add an l_2 loss on the predicted depth and the expected depth of the density field from the camera. When only a depth map is available, we use the N-dot-L shading image instead of the RGB image, which our method is robust to.

3.2. Lighting-conditioned diffusion

Our main goal is to provide explicit lighting control to a pre-trained diffusion model. Inspired by ControlNet [50], we train an additional module that provides control signals to the intermediate features of SD [34], as depicted in Fig. 3.

We use lighting information represented as direct shading maps as conditioning, which we concatenate to the normal map to provide geometric information to the model. Similarly to ControlNet [50], our control modules contain zero convolutions to introduce the control gradually.

During training, we keep SD [34] fixed and optimize only our control module consisting of a Residual Control Encoder and Decoder (RCE, RCD) and a Lighting Control network. We found that using the architecture of Control-Net [50] is prone to ignoring part of the control signal. Indeed as mentioned in the original paper, the controls tend to be picked up suddenly. We believe this might be due to low gradients early on as the encoder does not provide a meaningful signal to the control module. To avoid this issue, we develop a more stable encoder module, RCE, and next to the diffusion noise prediction loss we additionally supervise the training with an l_2 loss on the control reconstruction obtained with our RCD. RCE and RCD use residual blocks for more stable control flow and the reconstruction loss ensures
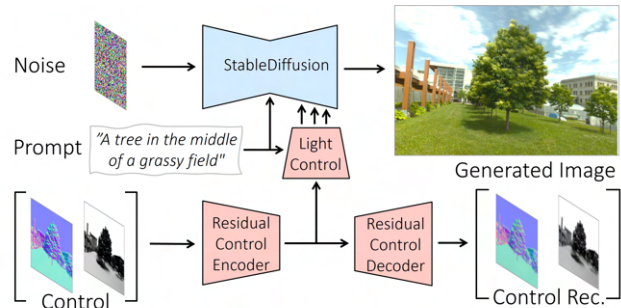

Figure 3. **Model Overview**. To generate lighting-controlled images, we train a light control module similar to [50], conditioned on normal and shading estimation. We use a custom Residual Control Encoder to encode the control signal for the ControlNet. Adding a Residual Control Decoder with a reconstruction loss ensures the full control signal is present in the encoded signal.

that the full control signal is provided to the light control module. During inference, we do not need the RCD.

3.3. Relighting Application

Besides controllable image generation, our lighting representation and dataset can also be employed for relighting applications. Relighting methods are usually trained on synthetic data leading to domain gap or on limited paired real data [40]. Adding relighting capability to pre-trained diffusion models opens up a novel way of utilizing pre-trained image priors. To achieve this, we propose to condition the generation on an input image and a target shading as opposed to normals and shading for the generation task.

Dataset. To avoid training on synthetic renderings leading to domain gap, we use predicted relit images (Fig. 4). Given cropped images and random lighting conditions, we use our shading estimation method to generate target coarse shading maps and predict relit images with OutCast [8].

Training. To avoid inheriting the artifacts of OutCast, we use the relit images as input and target the real image as output. This way, our output domain is real-world images, and our model is able to utilize the strong prior of SD [34].

3.4. Dataset

We use the Outdoor Laval Dataset [12], which consists of real-world HDR panoramas encoded as a latitude-longitude map. Given the full panorama, we determine the Sun's direction by detecting the brightest pixel in the panorama image and transforming it to an angular direction, which helps our shading estimation. Our dataset contains 51250 samples of LDR images and text prompts with corresponding estimated normal and shading maps, as shown in Fig. 4.

Image We crop 250 images of resolution 512×512 from each panorama. For each image, we use randomized camera parameters with varying field-of-view, elevation, and roll angles, as described in our supplemental. We normalize the images to have 0.5 mean intensity.

Depth and Normal As a first step in our dataset generation, we estimate the per-pixel surface normals of each image. To this end, we use the same depth estimator as OutCast [8], dubbed DepthNet. In summary, it is a segformer-based depth estimator [46] trained on the datasets proposed in [5, 17, 20, 25, 33, 42, 44]. The model was trained using the loss function proposed in [41]. We project this estimated depth to a point cloud using $x_i = \frac{1}{f} u_i z_i$, where u_i is the pixel's image coordinates, f is the focal length in pixels, and z_i is the estimated depth at pixel i. We perform the same operation on y_i to obtain $\vec{p}_i = [x_i, y_i, z_i]$. Finally, we obtain the per-pixel normal by computing the discrete derivative over the point cloud, as

$$\vec{n} = \frac{\partial \vec{p}}{\partial x} \times \frac{\partial \vec{p}}{\partial y} . \tag{1}$$

We experimentally compared this approach to directly estimating surface normals, and the former provided a more robust estimation. We hypothesize that the larger corpus of publicly available depth maps datasets yields a better depth estimator, explaining this improved performance.

At test time, we experimented with swapping our depth estimator with MiDaS [2, 32] and achieved similar image generation results, when a plausible depth map is obtained.

Shading A simple lighting representation is to employ the depth map as geometry and determine the N-dot-L shading. While this conceptually simple representation is straightforward to implement from a depth map alone, it does not consider cast shadows. Thus, we use our shading estimation method (Sec. 3.1) to obtain refined direct shading maps.

Prompt To maintain the textual capabilities of our model, we include text prompts for each sample in our dataset. We use BLIP-2 [19] to automatically caption the images.

4. Experiments

4.1. Image Synthesis

For image synthesis, our inputs consist of a shading map (Sec. 3.4) and optionally a normal map (Sec. 3.4). These
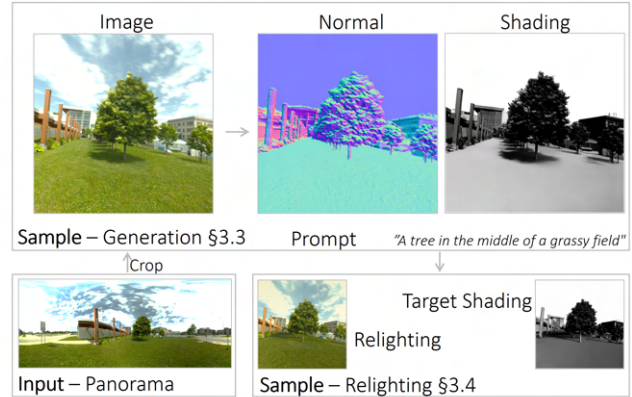

Figure 4. **Dataset Generation Pipeline**. We generate a dataset using the Outdoor Laval dataset [12]. We randomly crop images from the panoramas and automatically predict normal, shading, and caption (Sec. 3.4). For our relighting experiments (Sec. 3.3), we extend the dataset with relit images using OutCast [8].

maps can be estimated from either a guidance image or any text-to-image pipeline. . All evaluations are performed on real images from our test set or in-the-wild images, never seen in training. We want to emphasize that, to the best of our knowledge, our method is the first to achieve this degree of lighting control on diffusion-based generative imaging.

Training. We optimize our control module with the AdamW [21] optimizer for two epochs using a learning rate of $1e$-5 with a control reconstruction weight of 1.

Inference. We employ the DDPM sampler [11] for 1000 steps for quantitative and in-domain queries. For custom text prompts and styles, we use DDIM sampler [38] with 100 steps and a guidance scale of 7 and early control stopping at timestep 200 to avoid overruling the text guidance.

4.1.1 Lighting Consistency.

We first qualitatively evaluate our method's capability to produce the desired lighting across various text prompts, provided in Fig. 5. As can be seen, our method produces consistent, convincing shading across various styles following the target shading well.

We also evaluate the lighting consistency of our model with a user study on 42 participants shown in Tab. 1. The study contains images both from our test set and in the wild using the predicted text prompt from our dataset and also manually prepared ones. The users are presented with the input maps along with generated images of our method and SD [34] and are asked to answer three questions: 1) which image corresponds the best to the lighting input, 2) which image matches best the input text prompt, and 3) which image has the best overall quality. This user study reveals that our method not only follows the desired lighting well but is also preferred more in terms of image quality and textural alignment. Lighting is an essential part of the perceptual

Figure 5. Image Synthesis with Consistent Lighting. Our generated images feature consistent lighting aligned with the target shading for diverse text prompts.

Table 1. **Perceptual Image Generation Quality.** We perform a user study to assess the quality of the generated images. Since there exists no other method capable of explicit lighting control, we compare against SD [34]. The users are provided two images generated with SD [34] and our method using the same prompt. We report the perceptual quality regarding the image (I-PQ), lighting (L-PQ), and text alignment (T-PQ). Thanks to better lighting, our results are preferred in every aspect, not just in lighting.

	L-PQ ↑	I-PQ ↑	T-PQ ↑
SD	4.43	39.14	44.14
Ours	**95.57**	**60.86**	**55.86**

image quality. SD [34] is not enforced to produce physically consistent lighting, leading to perceptually degraded images compared to ours.

4.1.2 Lighting Controllability

In Fig. 6, we show examples of novel lighting on images from our test set. Our results correctly follow the user-defined lighting (insets) shown in the surface shading.

In Fig. 7, we show examples of novel image generation with controlled lighting on in-the-wild inputs. Specifically, the normal and shading maps of the top three rows come from images taken from the internet[1], while the two bottom rows were entirely generated from a text prompt using Stable Diffusion. We then ran our normal estimation and coarse shading estimation approach on each image entirely automatically. In this setup, each lighting was generated independently, without care for identity preservation; only the initial noise was fixed to mitigate discrepancies.

---

[1]We obtained the licenses for their use.

Table 2. **Relighting Evaluation.** We quantitatively compare our relighting method to OutCast on geometry (PSNR), image quality (FID, I-PQ) and lighting quality (L-PQ), where PQ refers to perceptual quality from our user study. The shadows are usually stronger for OutCast [8], leading to a slightly higher perceptual score. Our method achieves consistent relighting with more realistic image quality.

	PSNR ↑	FID ↓	L-PQ ↑	I-PQ ↑
OutCast	19.79 ± 4.39	71.08	**54.27**	36.47
Ours - w/o RCD	20.24 ± 3.29	76.28	-	-
Ours	**20.44** ± 3.34	**64.18**	45.74	**63.53**

4.2. Image Relighting

Our lighting representation and proposed dataset enable additional lighting-related applications, such as relighting. For this task, we use an input image and a target shading map as conditioning. We compare against our reimplementation of OutCast [8] using our shading estimation.

Training. We use our extended dataset (Sec. 3.3) and use the OutCast [8] relit image with the source shading as conditioning and the original image as target. We train this model for six epochs using the process described in Sec. 4.1.

Inference. We use the DDPM sampler with 1000 steps and produce the text prompts automatically using BLIP-2 [19].

Evaluation. We first validate our relighting model qualitatively on in the wild images compared against OutCast [8] in Fig. 8. OutCast provides very competitive results; however, being trained on synthetic data limits its generalization to real data. Regions originally in shadow notably suffer from noise amplification. In contrast, our diffusion-based model provides visually pleasing results.

We evaluate quantitatively in Tab. 2. We perform a cycle relighting experiment on our test set and predict the original real image from OutCast [8] relit images (PSNR). Our method already outperforms OutCast without our proposed
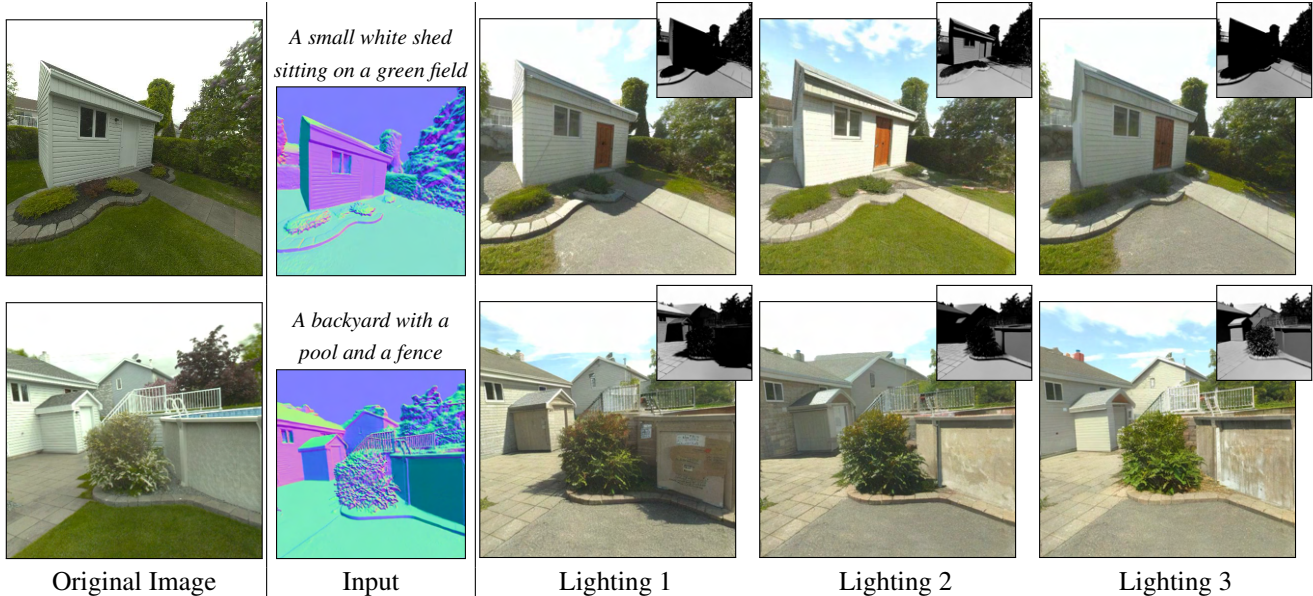
Figure 6. **In-Domain Image Synthesis with Controllable Lighting**. We can synthesize images under various lighting conditions.

RCD module (Sec. 3.2), thanks to the diffusion prior. In addition, our full method produces more natural images with the best FID score [10]. We further evaluate the perceptual quality (PQ) with a user study performed by 17 people. The users were asked to evaluate the lighting consistency (L-PQ) and the overall image quality and realism (I-PQ) between our method and OutCast. OutCast usually provides stronger shadows that are perceptually slightly more consistent. However, aligned with the FID, our method produces significantly more realistic results according to the users.

4.3. Ablations

4.3.1 Image Synthesis

Does the model need cast shadows? Our key design choice is to provide information about cast shadows to the model. We argue that simpler lighting representation is insufficient because the latent space of a pre-trained diffusion model does not encode consistent global illumination. We qualitatively compare against a similar but simpler setup in Fig. 10, where the conditioning is an N-dot-L shading map without any cast shadow. Notably, our model can infer the overall lighting also with N-dot-L shading but fails to generate realistic shadows. In contrast, using direct shading offers much more appealing results with fine-grained shadow control. We provide more examples in our supplemental.

We provide quantitative results in Tab. 3. We consider the estimated shading quality (L-PSNR) and the angular error between dominant light directions (L-AE). Using direct shading outperforms N-dot-L with a high margin.
Does the model need normals? Although most of the

Table 3. **Control Consistency.** We estimate the shading (L) and normal (N) of generated images on our test set and compare them against the control signal in image space (PSNR) and in angular error measured in degrees (AE). Conditioning on our direct shading (DS) achieves the best lighting quality; however, it does not ensure consistent normals (Fig. 11). Providing normals to the model helps with minimal cost of the lighting quality.

	L-PSNR ↑	L-AE ↓	N-PSNR ↑	N-AE ↓
N-dot-L Shading	6.43 ± 2.20	37.23 ± 23.79	16.45 ± 2.53	21.76 ± 9.96
Direct Shading (DS)	**13.04** ± 3.57	**27.30** ± 19.13	17.30 ± 2.75	18.73 ± 9.54
DS + Normals (Ours)	12.69 ± 3.52	28.59 ± 20.46	**17.47** ± 2.72	**18.28** ± 9.28

geometry can be inferred from our direct shading map, shadow regions do not provide any signal. When the incident light is away from a surface normal $\cos(\vec{n} \cdot \vec{l}) \leq 0$ (attached shadow) or the light is occluded by some geometry (cast shadow) results in a uniform region of null values. Without additional geometric information, the model generates random geometric detail in those regions. We showcase this in Fig. 11, where the model without normal generates a flat wall devoid of features in the shadow region, while our method correctly generates the expected door and windows.

We quantitatively evaluate the effect of normal conditioning on our test set in Tab. 3. Using normals improves the normal consistency in the shadow regions.

4.3.2 Image Relighting

Does our architecture help identity preservation? Identity preservation is a crucial aspect of image relighting. Unfortunately, we have witnessed that diffusion-based image

Figure 7. **Out-of-Domain Image Synthesis with Controllable Lighting**. Our method learns to control the generation process yet maintains the prior of SD [34]. We show various scenes and styles under changing lighting conditions. The first three images were obtained with estimated normal of real-world images, while for the last two, we used images generated with SDXL [30]. See supplemental for details.
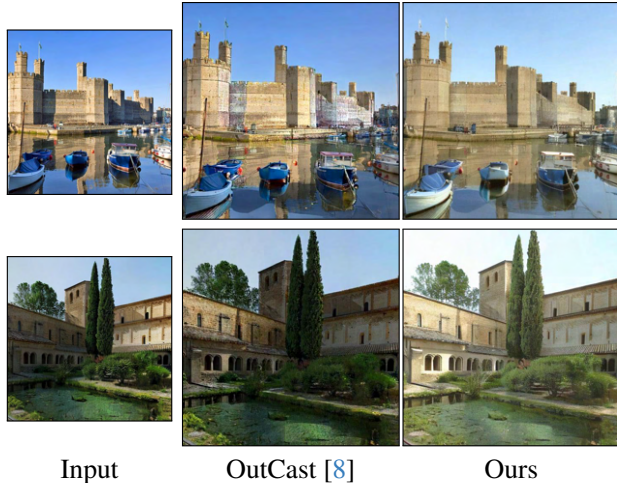
Figure 8. **Relighting of Real-World Images**. We train a relighting network and evaluate it on real-world images. Utilizing the diffusion prior helps the generalization to real samples, especially for shading disambiguation.
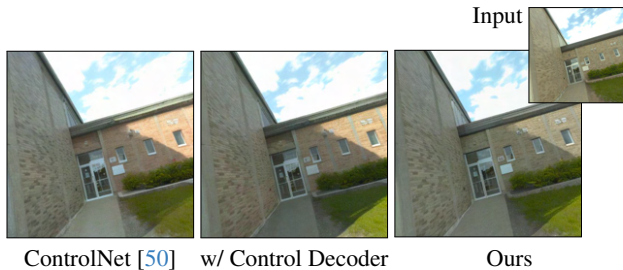

Figure 9. **Identity Preservation**. We ablate the effect of our control architecture on relighting. ControlNet [50] - left - is prone to ignoring part of the control signal, the wall turns reddish and the shadow gets softened. Our Control Decoder - middle - with control reconstruction loss helps. Our full residual architecture - right - takes another step and achieves high consistency.

editing generally exhibits issues in identity preservation, especially in reproducing colors. Directly training ControlNet [50] on our task produces changes in wall tint, for example, as shown in Fig. 9. We hypothesize that information pertaining to identity is lost in the encoder. which ensures that the feature map injected into the denoising U-Net keeps information to maintain the control signal.

Tab. 2 quantitatively showcases the importance of our RCD (Sec. 3.2), where the image quality increases to an FID of 64.18 when our method is used without it.

4.4. Limitations and Future Work

Our work assumes directional lighting, which is suitable for outdoor scenes. However, our shading estimation method enables tracing rays in arbitrary directions. Adapting our method to point and other light sources is an exciting avenue for future research. Furthermore, our shading estimation requires lighting direction for the best results. Combining
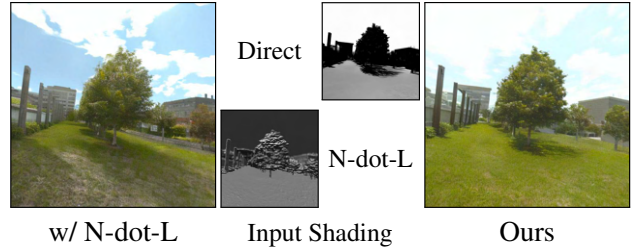

Figure 10. **Effect of Lighting Representation**. We show that cast shadows provide essential information for the generation process. We compare our method against a simple N-dot-L shading conditioning, which provides only coarse lighting information to the model, leading to inconsistent lighting with less control.
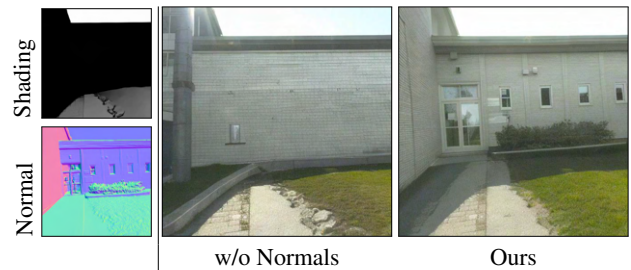

Figure 11. **Effect of Normal Conditioning**. Without normal conditioning, it is impossible for the model to infer geometry in the shadowed regions.

our method with a robust lighting estimation would allow training on much larger datasets.

5. Conclusion

Recent diffusion-based generative imaging techniques have shown impressive text-to-image capabilities, producing breathtaking images on a whim. However, their controllability is limited, and adjusting important details such as lighting requires careful prompt engineering. In this work, we present a novel approach to explicitly control the illumination of images generated by a diffusion model. Our approach uses our direct shading representation, which contains both shading and shadow information. The shading map can be automatically computed from an existing picture or a generated image. Our method achieves high-quality results compared to existing methods while maintaining user-defined lighting. We believe that our method paves the way to increase the editability of diffusion-based generative imaging approaches.

References

[1] Jonathan T Barron and Jitendra Malik. Shape, illumination, and reflectance from shading. *IEEE transactions on pattern analysis and machine intelligence*, 37(8):1670–1687, 2014. 2

[2] Reiner Birkl, Diana Wofk, and Matthias Müller. Midas v3. 1–a model zoo for robust monocular relative depth estimation. *arXiv preprint arXiv:2307.14460*, 2023. 4

[3] Riccardo Corvi, Davide Cozzolino, Giada Zingarini, Giovanni Poggi, Koki Nagano, and Luisa Verdoliva. On the detection of synthetic images generated by diffusion models. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023. 1

[4] Paul Debevec, Tim Hawkins, Chris Tchou, Haarm-Pieter Duiker, Westley Sarokin, and Mark Sagar. Acquiring the reflectance field of a human face. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, pages 145–156, 2000. 2

[5] Rahul Garg, Neal Wadhwa, Sameer Ansari, and Jonathan T Barron. Learning single camera depth estimation using dualpixels. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7628–7637, 2019. 4

[6] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2414–2423, 2016. 2

[7] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014. 2

[8] David Griffiths, Tobias Ritschel, and Julien Philip. Outcast: Outdoor single-image relighting with cast shadows. In *Computer Graphics Forum*, pages 179–193. Wiley Online Library, 2022. 2, 3, 4, 5, 8

[9] 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

[10] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 6

[11] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 2, 4

[12] Yannick Hold-Geoffroy, Akshaya Athawale, and Jean-François Lalonde. Deep sky modeling for single image outdoor lighting estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6927–6935, 2019. 4, 1

[13] Lianghua Huang, Di Chen, Yu Liu, Yujun Shen, Deli Zhao, and Jingren Zhou. Composer: Creative and controllable image synthesis with composable conditions. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, pages 13753–13773. PMLR, 2023. 1

[14] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei 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. 2

[15] Yongcheng Jing, Yezhou Yang, Zunlei Feng, Jingwen Ye, Yizhou Yu, and Mingli Song. Neural style transfer: A review. *IEEE transactions on visualization and computer graphics*, 26(11):3365–3385, 2019. 2

[16] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 2

[17] Anastasiia Kornilova, Marsel Faizullin, Konstantin Pakulev, Andrey Sadkov, Denis Kukushkin, Azat Akhmetyanov, Timur Akhtyamov, Hekmat Taherinejad, and Gonzalo Ferrer. Smartportraits: Depth powered handheld smartphone dataset of human portraits for state estimation, reconstruction and synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21318–21329, 2022. 4

[18] Max Ku, Tianle Li, Kai Zhang, Yujie Lu, Xingyu Fu, Wenwen Zhuang, and Wenhu Chen. Imagenhub: Standardizing the evaluation of conditional image generation models. *arXiv preprint arXiv:2310.01596*, 2023. 2

[19] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023. 4, 5

[20] Zhengqi Li and Noah Snavely. Megadepth: Learning single-view depth prediction from internet photos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2041–2050, 2018. 4

[21] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 4

[22] Fujun Luan, Sylvain Paris, Eli Shechtman, and Kavita Bala. Deep photo style transfer. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4990–4998, 2017. 2

[23] Chong Mou, Xintao Wang, Jiechong Song, Ying Shan, and Jian Zhang. Dragondiffusion: Enabling drag-style manipulation on diffusion models. *CoRR*, abs/2307.02421, 2023. 1

[24] Chong Mou, Xintao Wang, Liangbin Xie, Jian Zhang, Zhongang Qi, Ying Shan, and Xiaohu Qie. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. *arXiv preprint arXiv:2302.08453*, 2023. 1, 2

[25] Pushmeet Kohli Nathan Silberman, Derek Hoiem and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012. 4

[26] Baptiste Nicolet, Julien Philip, and George Drettakis. Repurposing a relighting network for realistic compositions of captured scenes. In *Symposium on Interactive 3D Graphics and Games*, pages 1–9, 2020. 2

[27] Julien Philip, Michaël Gharbi, Tinghui Zhou, Alexei A Efros, and George Drettakis. Multi-view relighting using a geometry-aware network. *ACM Trans. Graph.*, 38(4):78–1, 2019.

[28] Julien Philip, Sébastien Morgenthaler, Michaël Gharbi, and George Drettakis. Free-viewpoint indoor neural relighting from multi-view stereo. *ACM Transactions on Graphics (TOG)*, 40(5):1–18, 2021. 2

[29] Ryan Po, Wang Yifan, Vladislav Golyanik, Kfir Aberman, Jonathan T. Barron, Amit H. Bermano, Eric Ryan Chan, Tali Dekel, Aleksander Holynski, Angjoo Kanazawa, C. Karen Liu, Lingjie Liu, Ben Mildenhall, Matthias Nießner, Björn Ommer, Christian Theobalt, Peter Wonka, and Gordon Wetzstein. State of the art on diffusion models for visual computing. *CoRR*, abs/2310.07204, 2023. 2

[30] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: improving latent diffusion models for high-resolution image synthesis. *CoRR*, abs/2307.01952, 2023. 7, 1

[31] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015. 2

[32] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE transactions on pattern analysis and machine intelligence*, 44(3):1623–1637, 2020. 4

[33] Mike Roberts, Jason Ramapuram, Anurag Ranjan, Atulit Kumar, Miguel Angel Bautista, Nathan Paczan, Russ Webb, and Joshua M Susskind. Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10912–10922, 2021. 4

[34] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 2, 3, 4, 5, 7, 1

[35] Viktor Rudnev, Mohamed Elgharib, William Smith, Lingjie Liu, Vladislav Golyanik, and Christian Theobalt. Nerf for outdoor scene relighting. In *European Conference on Computer Vision*, pages 615–631. Springer, 2022. 2

[36] Soumyadip Sengupta, Brian Curless, Ira Kemelmacher-Shlizerman, and Steven M Seitz. A light stage on every desk. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2420–2429, 2021. 2

[37] Yujun Shi, Chuhui Xue, Jiachun Pan, Wenqing Zhang, Vincent Y. F. Tan, and Song Bai. Dragdiffusion: Harnessing diffusion models for interactive point-based image editing. *CoRR*, abs/2306.14435, 2023. 1

[38] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 1, 2, 4

[39] Tiancheng Sun, Jonathan T Barron, Yun-Ta Tsai, Zexiang Xu, Xueming Yu, Graham Fyffe, Christoph Rhemann, Jay Busch, Paul Debevec, and Ravi Ramamoorthi. Single image portrait relighting. *ACM Transactions on Graphics (TOG)*, 38(4):1–12, 2019. 2

[40] Ayush Tewari, Ohad Fried, Justus Thies, Vincent Sitzmann, Stephen Lombardi, Kalyan Sunkavalli, Ricardo Martin-Brualla, Tomas Simon, Jason Saragih, Matthias Nießner, et al. State of the art on neural rendering. In *Computer Graphics Forum*, pages 701–727. Wiley Online Library, 2020. 2, 3

[41] Benjamin Ummenhofer, Huizhong Zhou, Jonas Uhrig, Nikolaus Mayer, Eddy Ilg, Alexey Dosovitskiy, and Thomas Brox. Demon: Depth and motion network for learning monocular stereo. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5038–5047, 2017. 4

[42] Igor Vasiljevic, Nick Kolkin, Shanyi Zhang, Ruotian Luo, Haochen Wang, Falcon Z Dai, Andrea F Daniele, Mohammadreza Mostajabi, Steven Basart, Matthew R Walter, et al. Diode: A dense indoor and outdoor depth dataset. *arXiv preprint arXiv:1908.00463*, 2019. 4

[43] Bram Wallace, Akash Gokul, and Nikhil Naik. EDICT: exact diffusion inversion via coupled transformations. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 22532–22541. IEEE, 2023. 1

[44] Qiang Wang, Shizhen Zheng, Qingsong Yan, Fei Deng, Kaiyong Zhao, and Xiaowen Chu. Irs: A large naturalistic indoor robotics stereo dataset to train deep models for disparity and surface normal estimation. In *2021 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2021. 4

[45] Zian Wang, Tianchang Shen, Jun Gao, Shengyu Huang, Jacob Munkberg, Jon Hasselgren, Zan Gojcic, Wenzheng Chen, and Sanja Fidler. Neural fields meet explicit geometric representations for inverse rendering of urban scenes. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[46] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*, 34:12077–12090, 2021. 4

[47] Zexiang Xu, Kalyan Sunkavalli, Sunil Hadap, and Ravi Ramamoorthi. Deep image-based relighting from optimal sparse samples. *ACM Transactions on Graphics (ToG)*, 37 (4):1–13, 2018. 2

[48] Ye Yu, Abhimitra Meka, Mohamed Elgharib, Hans-Peter Seidel, Christian Theobalt, and William AP Smith. Self-supervised outdoor scene relighting. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXII 16*, pages 84–101. Springer, 2020. 2

[49] Guanqi Zhan, Chuanxia Zheng, Weidi Xie, and Andrew Zisserman. What does stable diffusion know about the 3d scene? *arXiv preprint arXiv:2310.06836*, 2023. 2

[50] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models, 2023. 1, 2, 3, 8

[51] Hao Zhou, Sunil Hadap, Kalyan Sunkavalli, and David W Jacobs. Deep single-image portrait relighting. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7194–7202, 2019. 2