LightIt: Illumination Modeling and Control for Diffusion Models

Supplementary Material

In this supplementary material, first, we provide additional details on our method in Appendix A, and on our experiments in Appendix B. Finally, we show additional results in Appendix C.

A. Method Details

We provide additional details about our method in the following sections.

A.1. Dataset Generation

Image. Our dataset is generated from the Outdoor Laval dataset [12] consisting of 205 HDR panorama images. Having access to the full panorama image gives us beneficial information about the dominant light direction.

We render 250 images from virtual cameras for every panorama. We use varying intrinsic and extrinsic parameters for the views with the constraints shown in Tab. 4. We normalize the images to have a mean intensity of 0.5. We convert the images to LDR format by transforming to sRGB space using gamma correction ($\gamma = 2.2$) and clipping to the $[0, 1)$ range.

A.2. Lighting Control

Our full control module consists of three main modules: Residual Control Encoder (RCE), Lighting Control (LC) used during both training and inference; Residual Control Decoder (RCD) used only during training. RCE has approximately 2.9M, LC 363M, and RCD 1.3M parameters.

RCE. Our RCE consists of 7 residual blocks and maps from $512 \times 512 \times 3$ to $64 \times 64 \times 320$. The residual blocks follow the architecture of the ones in the diffusion UNet of [34]. Following [50], we use zero convolution at the beginning of our RCE.

RCD. During training, we use a decoder to reconstruct the control signal from the latent representation and use a reconstruction supervision to ensure most of the signal is preserved. Our RCD significantly improves the control consistency. Similar to our RCE, RCD consists of 7 residual blocks and uses the same architecture as RCE, just in transposed order.

LC. Following [50], our LC uses the same architecture as the encoder of the UNet of [34]. LC takes the encoded control signal and returns the intermediate and encoded diffusion features to control the diffusion process [50].

Table 4. **Image Cropping Parameters.** We crop images from the [12] dataset using the following parameters in degrees.

	Min	Max	Distribution
Vertical FOV	30	110	Uniform
Azimuth	0	360	Uniform
Elevation	-22.5	22.5	Triangular
Roll	-22.5	22.5	Triangular

B. Experiment Details

Lighting Consistency (Sec. 4.1.1). We conduct a user study to perceptually evaluate the real-world predictions of our generated images. We provide all the samples together with the results in the `generation_results.html` of our supplementary material.

Lighting Controllability (Sec. 4.1.2). For the samples shown in Figure 7, we used in-the-wild images as well as generated images using SDXL [30]. We show the original images, the estimated normals, and the used text prompts in Fig. 13. Over the different lighting conditions, we fixed the seed. To obtain the SDXL-generated [30] samples, we used the same prompt as for our generation.

Relighting (Sec. 4.2). We conduct a user study to perceptually evaluate the real-world predictions of our relighting application. We provide all the samples together with the results in the `relighting_results.html` of our supplementary material.

C. Additional Results

Effect of Lighting Representation (Sec. 4.3.1). Our method uses direct shading, which contains information about cast shadows. Since SD [34] does not model light transport, cast shadows provide valuable information to the model. We show additional samples to our ablation (Fig. 10) in Fig. 12.
Control robustness (Sec. 3.4).

Figure 12. **Effect of Lighting Representation**. We show additional results to our ablation about the effect of cast shadows (Fig. 10).

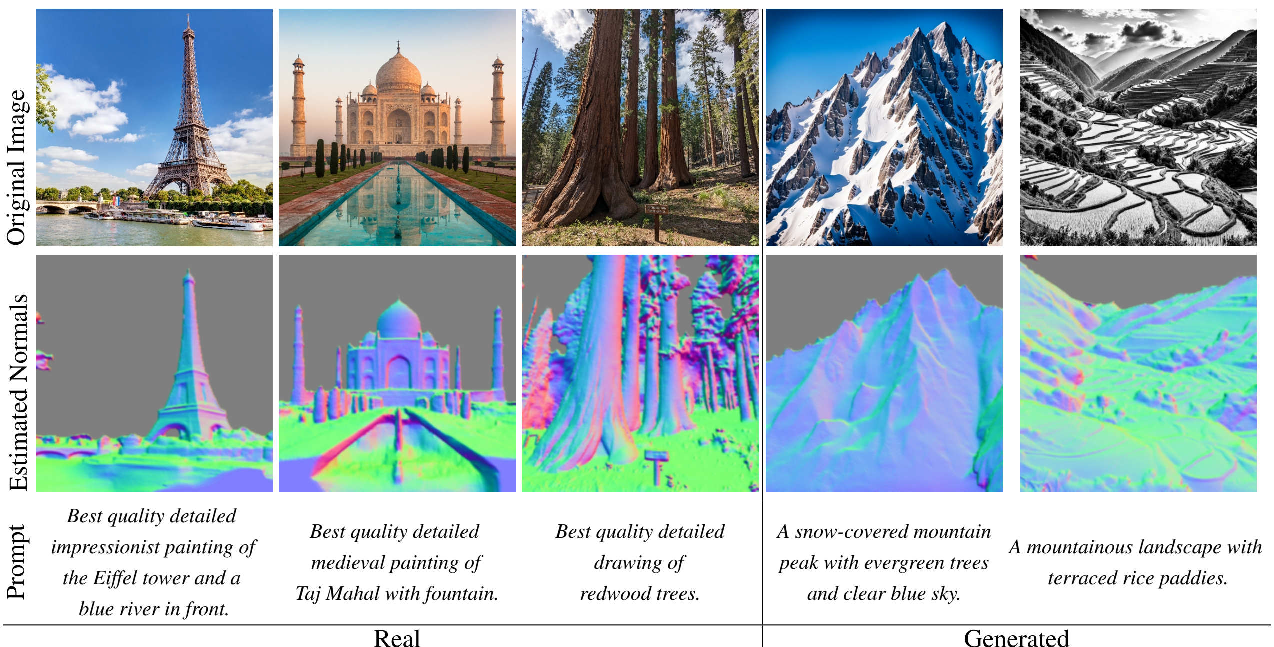

Figure 13. **Out-of-Domain Image Synthesis Details**. We show the original input image together with the estimated normals and text prompts used to generate our images in (Fig. 7).

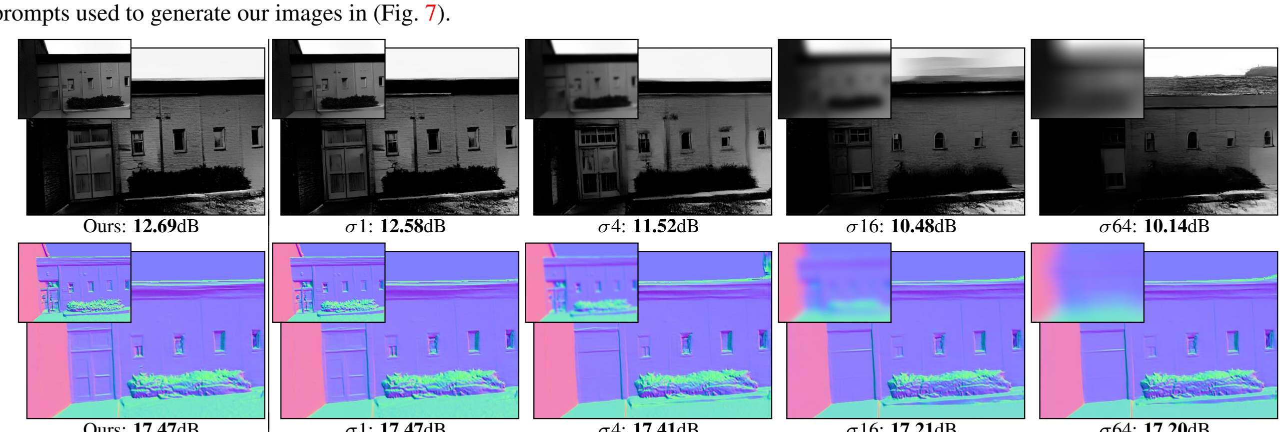

Figure 14. **Control robustness**. Test set average for original control reconstruction of our generated samples with increasingly blurred input (top left insets).