

A Diffusion Approach to Radiance Field Relighting using Multi-Illumination Synthesis

Y. Poirier-Ginter^{1,2}, A. Gauthier¹, J. Philip³, J.-F. Lalonde² and G. Drettakis¹
¹Inria, Université Côte d'Azur, France; ²Université Laval, Canada; ³Adobe Research, United Kingdom

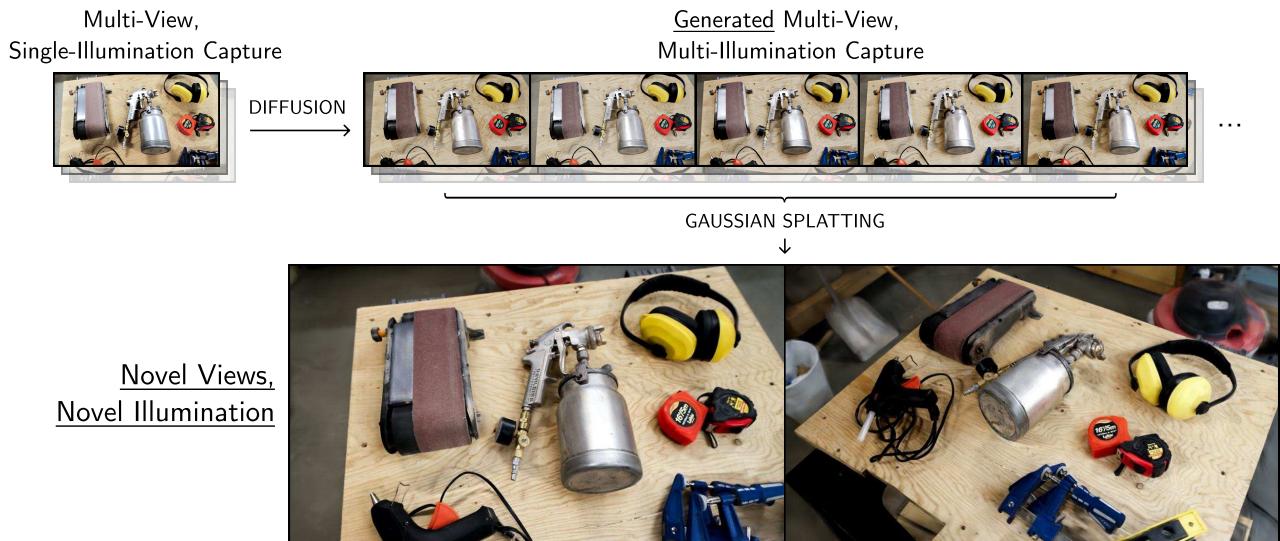


Figure 1: Our method produces relightable radiance fields directly from single-illumination multi-view dataset, by using priors from generative data in the place of an actual multi-illumination capture.

Abstract

Relighting radiance fields is severely underconstrained for multi-view data, which is most often captured under a single illumination condition. It is especially hard for full scenes containing multiple objects. We introduce a method to create relightable radiance fields using such single-illumination data by exploiting priors extracted from 2D image diffusion models. We first fine-tune a 2D diffusion model on a multi-illumination dataset conditioned by light direction, allowing us to augment a single-illumination capture into a realistic – but possibly inconsistent – multi-illumination dataset from directly defined light directions. We use this augmented data to create a relightable radiance field represented by 3D Gaussian splats. To allow direct control of light direction for low-frequency lighting, we represent appearance with a multi-layer perceptron parameterized on light direction. To enforce multi-view consistency and overcome inaccuracies we optimize a per-image auxiliary feature vector. We show results on synthetic and real multi-view data under single illumination, demonstrating that our method successfully exploits 2D diffusion model priors to allow realistic 3D relighting for complete scenes.

Keywords: NeRF, Radiance Field, Relighting

1. Introduction

Radiance fields have recently revolutionized 3D scene capture from images [MST*20]. Such captures typically involve a multi-view set of photographs taken under the *same* lighting conditions. Relighting such radiance fields is hard since lighting and material properties are entangled (e.g., is this a shadow or simply a darker color?) and the inverse problem ill-posed.

One approach to overcome this difficulty is to capture a multi-illumination dataset which better conditions the inverse problem but comes at the cost of a heavy capture setup [DHT*00]. Another option is to use *priors*, which is typically done by training a neural network on synthetic data to predict intrinsic properties or relit images. However, creating sufficiently large, varied and photorealistic 3D scenes is both challenging and time-consuming. As such, methods relying on these—or simpler—priors often demonstrate

results on isolated masked objects [BBJ^{*}21], or make simplifying assumptions such as distant environment lighting [BBJ^{*}21, ZSD^{*}21]. Other methods have handled more complex illumination models, including full scenes [PMGD21, PGZ^{*}19], but can be limited in the complexity of the geometry and materials that must reconstruct well. Finally, methods that depend on accurate estimates of surface normals [JLX^{*}23, GGL^{*}23] often produce limited levels of realism when relighting.

At the other end of the spectrum, diffusion models (DMs, e.g., [RBL^{*}22]), trained on billions of natural images, have shown exceptional abilities to capture real image distribution priors and can synthesize complex lighting effects. While recent progress shows they can be controlled in various ways [ZRA23], extracting lighting-specific priors from these models, especially for full 3D scenes, has not yet been demonstrated.

In this paper, we build on these observations and present a new method that demonstrates that it is possible to create relightable radiance fields for complete scenes from single low-frequency lighting condition captures by exploiting 2D diffusion model priors. We first propose to fine-tune a pre-trained DM conditioned on the dominant light source direction. For this, we leverage a dataset of images with many lighting conditions of the same scene [MGAD19], which enables the DM to produce relit versions of an image with explicit control over the dominant lighting direction. We use this 2D relighting network to augment any standard multi-view dataset taken under single lighting by generating multiple relit versions of each image, effectively transforming it into a multi-illumination dataset. Given this augmented dataset, we train a new *relightable radiance field* with direct control on lighting direction, which in turn enables realistic interactive relighting of full scenes with lighting and camera view control in real time for low-frequency lighting. We build on 3D Gaussian Splatting (3DGS) [KKLD23] introduces an explicit, primitive-based representation of radiance fields. The anisotropic nature of the 3D Gaussians allows the efficient representation of fine detail, and the fast GPU-accelerated rasterization used allows real-time rendering. We use 3DGS to represent radiance fields mainly for performance, but any other radiance representation, e.g., [CXG^{*}22], could be used instead. Radiance fields are most commonly used in the context of *single-light condition* captures, i.e., the images are all captured under the same lighting. As a result, there is no direct way to change the lighting of captured scenes, severely restricting the utility of radiance fields compared to traditional 3D graphics assets. Our method uses diffusion models to simulate multi-light conditions from a single-light capture thus allowing the relighting of radiance fields.

In summary, our contributions are:

- A new 2D relighting neural network with direct control on lighting direction, created by fine-tuning a DM with multi-lighting data.
- A method to augment single-lighting multi-view capture to an approximate multi-lighting dataset, by exploiting the 2D relighting network.
- An interactive relightable radiance field that provides direct control on lighting direction, and corrects for inconsistencies in the neural relighting.

We demonstrate our solution on synthetic and real indoor scenes, showing that it provides realistic relighting of multi-view datasets captured under a single lighting condition in real time.

2. Related Work

Our method proposes a relightable radiance field. We review work on radiance fields and their relightable variants, and discuss diffusion models and fine-tuning methods we build on.

2.1. Radiance Fields

Radiance field methods have revolutionized 3D scene capture using multi-view datasets (photos or video) as input. In particular, Neural Radiance Fields (NeRFs) [MST^{*}20] learn to synthesize novel views of a given scene by regressing its radiance from a set of input images (multiple photos or videos of a 3D scene). Structure from motion [Ull79, SF16] is used to estimate the camera poses for all images and rays are cast through the center of all pixels. A multi-layer perceptron (MLP) c_θ parameterized by 3D position and view direction is used to represent the radiance and opacity of the scene. The optimization objective is simply the mean squared error:

$$\mathcal{L}_{\text{NeRF}} = \mathbb{E}_{o,d,c^*} \left[\|c_\theta(o, d) - c^*\|_2^2 \right],$$

where o is a ray's origin, d its direction, and c^* the target RGB color value of its corresponding pixel. The predicted color for that pixel is obtained by integrating a color field c_θ weighted by a density field σ_θ following the equation of volume rendering. The original NeRF was slow to train and to render; A vast number of methods [TTM^{*}22] have been proposed to improve the original technique, e.g., acceleration structures [MESK22], antialiasing [BMT^{*}21], handling larger scenes [BMV^{*}22] etc. Recently, 3D Gaussian Splatting (3DGS) [KKLD23] introduces an explicit, primitive-based representation of radiance fields. The anisotropic nature of the 3D Gaussians allows the efficient representation of fine detail, and the fast GPU-accelerated rasterization used allows real-time rendering. We use 3DGS to represent radiance fields mainly for performance, but any other radiance representation, e.g., [CXG^{*}22], could be used instead. Radiance fields are most commonly used in the context of *single-light condition* captures, i.e., the images are all captured under the same lighting. As a result, there is no direct way to change the lighting of captured scenes, severely restricting the utility of radiance fields compared to traditional 3D graphics assets. Our method uses diffusion models to simulate multi-light conditions from a single-light capture thus allowing the relighting of radiance fields.

2.2. Single Image Relighting

Single image relighting approaches have mostly been restricted to human faces, with the most recent methods using generative priors [WZL^{*}08, SYH^{*}17, SKCJ18, FRV^{*}23, PTS23, PLMZ23]. Recently, human body relighting has also been studied [KE19, LSY^{*}21] as their structure also allows for the design of effective priors. Because they are much less constrained, relighting generic scenes from a single image is a much harder problem that has eluded researchers until recently. While some approaches have focused on specific relighting effects such as shadows [LLZ^{*}20, SLZ^{*}22, SZB21, VZG^{*}23], they are applicable solely for the purpose of object compositing. Full scene relighting has been explored by Murmann et al. [MGAD19], who present a dataset of real indoor scenes lit by multiple lighting directions. They show that training a U-net on their dataset allows for full scene relighting—in this work, we also leverage their dataset but train a more powerful ControlNet [ZRA23] for the relighting task. Other works include [TCE^{*}21] who focus on sky relighting and [LSB^{*}22, ZLZ^{*}22] which leverage image-to-image translation. Methods for outdoor scenes have also been proposed [YME^{*}20, LGZ^{*}20, YS22]. Of note, SIMBAR [ZTS^{*}22] and

OutCast [GRP22] produces realistic, user-controllable, hard, cast shadows from the sun. In contrast, we focus on indoor scenes which often exhibit soft shadows and more complex lighting effects. Finally, the concurrent work of Zeng et al. [ZDP^{*}24] uses diffusion models to relight isolated objects using environment maps. In contrast to these solutions, we focus on cluttered indoor scenes which often exhibit soft shadows and more complex lighting effects.

2.3. Multi-view Relighting

While single-view methods produce good results on restricted datasets such as faces, they are often limited by the lack of accurate geometry, required to simulate light transport. To this point, multi-view data can provide a more accurate and complete geometric reconstruction. For example Philip et al. [PMGD21, PGZ^{*}19] build on multi-view stereo (MVS) reconstruction of the scene, and learn a prior from synthetic data rendered under multiple lighting conditions. Despite correcting for many of the reconstruction artifacts, these methods are restricted by the quality of MVS reconstruction. Nimier et al. [NDDJK21] also present a scene-scale solution but require a complex pipeline that optimizes in texture space. Gao et al. [GCD^{*}20] use a rough proxy and neural textures to allow object relighting.

More recently radiance fields have also been used as a geometric representation for relighting. Most methods work on the simple case of a single isolated object while we target larger scenes. Such methods typically assume lighting to be distant, often provided as an environment map. NeRFactor [ZSD^{*}21] uses a Bi-Directional Reflectance Distribution Function (BRDF) prior from measurements and estimates normals and visibility, while NERV [SDZ^{*}21] and Zhang et al. [ZSH^{*}22] predicts visibility to allow indirect illumination estimation. NERD [BBJ^{*}21], PhySG [ZLW^{*}21], and DE-NeRF [WSLG23] use efficient physically-based material and lighting models to decompose a radiance field into spatially varying BRDFs, while Neural-PIL [BBJ^{*}21] learns the illumination integration and low-dimensional BRDF priors using auto-encoders. TensorIR [JLX^{*}23] uses a mixed radiance and physics-based formulation to recover intrinsic properties. NeRO [LWL^{*}23] focuses on specular objects showing very promising results. Relightable Gaussians [GGL^{*}23] use the more recent 3D Gaussian representation along with ray-tracing to estimate properties of objects. GS-IR [LZF^{*}24], GaussianShader [JTF^{*}24] and GIR [SWW^{*}23] also build on 3D Gaussian splatting, proposing different approaches to estimate more reliable normals while approximating visibility and indirect illumination; these work well for isolated objects under distant lighting. However, these methods struggle with more complex scene-scale input and near-field illumination but can work or be adapted to both single and multi-illumination input data.

Feeding multi-view multi-illumination data to a relightable radiance field indeed enables better relighting but at the cost of highly-controlled capture conditions [XZC^{*}23, ZCD^{*}23, TDMS^{*}23] or an extended dataset of unconstrained illuminations [BEK^{*}22, LGF^{*}22]. In our method, we use a Diffusion Model to simulate multi-illumination data, lifting the capture constraints while benefiting from the lighting variations. Another body of work [MHS^{*}22, HHM22, YZL^{*}22, ZYL^{*}23, LWC^{*}23, LCL^{*}23] achieve object or scene relighting from multi-view images by extracting traditional

3D assets (meshes and SVBRDFs) and applying physically-based rendering algorithms. IBL-NeRF [CKK23] allows for scene-scale material estimation but bakes the illumination into a prefiltered light-field which prevents relighting. Recently, NeRF-OSR [RES^{*}22], I²-SDF [ZHY^{*}23], and Wang et al. [WSG^{*}23] focused on scene scale, single illumination relighting scenes using both implicit and explicit representations. While they can achieve reasonable results, they often lack overall realism, exhibiting bumpy or overly smooth shading during relighting. In contrast, our use of diffusion priors provides realistic-looking output.

2.4. Diffusion Models

Diffusion Models (DMs) [SDWMG15, HJA20] made it possible to train generative models on diverse, high-resolution datasets of billions of images. These models learn to invert a forward diffusion process that gradually transforms images into isotropic Gaussian noise, by adding random Gaussian noise $\epsilon_t \sim \mathcal{N}(0, I)$ to an image in T steps. DMs train a neural network g_ϕ with parameters ϕ to learn to denoise with the objective:

$$\mathcal{L}_{\text{Diffusion}} = \mathbb{E}_{x, \epsilon, 1 \leq t \leq T} \left[\|g_\phi(x_t | t) - y_t\|_2^2 \right],$$

in which target y_t is often set to ϵ . After training, sampling can be performed step-by-step, by predicting x_{t-1} from x_t for each timestep t which is expensive since T can be high (e.g., 1000); faster alternatives include deterministic DDIM [SME20] sampling, that can perform sampling of comparable quality with fewer steps (i.e., 10-50× larger steps). Stable Diffusion [RBL^{*}22] performs denoising in a lower-dimensional latent space, by first training a variational encoder to compress images; for instance, in Stable Diffusion XL [PEL^{*}23], images are mapped to a latent space of size $\mathbb{R}^{128 \times 128 \times 4}$. In a pre-pass, the dataset is compressed using this autoencoder, and a text-conditioned diffusion model is then trained directly in this latent space.

Diffusion models have an impressive capacity to synthesize highly realistic images, typically conditioned on text prompts. The power of DMs lies in the fact that the billions of images used for training contain an extremely rich representation of the visual world. However, extracting the required information for specific tasks, without incurring the (unrealistic) cost of retraining DMs is not straightforward. A set of recent methods show that it is possible to fine-tune DMs with a typically much shorter training process to perform specific tasks (e.g., [GAA^{*}23, RLJ^{*}23]). A notable example is ControlNet [ZRA23] which proposed an efficient method for fine-tuning Stable Diffusion with added conditioning. In particular, they demonstrated conditional generation from depth, Canny edges, etc., with and without text prompts; We will build on this solution for our 2D relighting method.

In a similar spirit, there has been significant evidence in recent years that latent spaces of generative models encode material information [BMHF23, BF24]. Recent work shows the potential to fine-tune DMs to allow direct material editing [SJL^{*}23]. Nonetheless, we are unaware of published methods that use DM fine-tuning to perform realistic relighting of full and cluttered scenes.

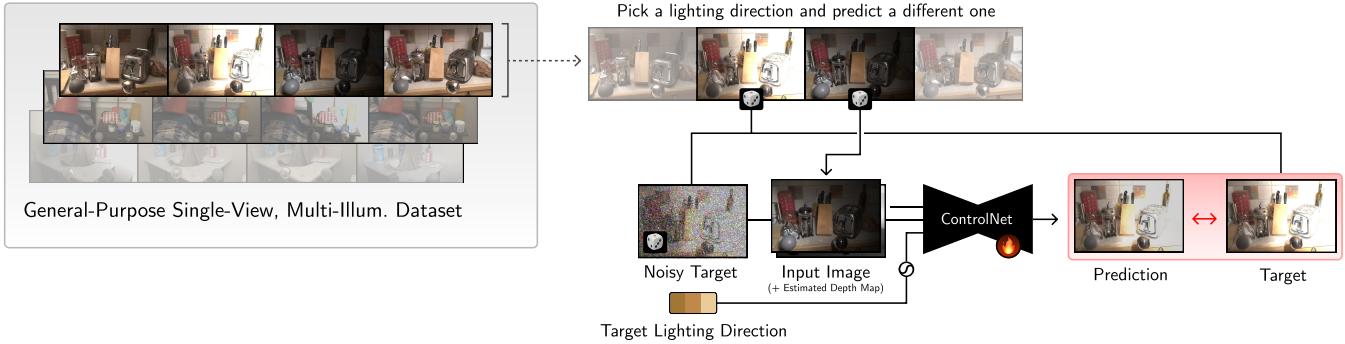


Figure 2: We use the single-view, multi-illumination dataset of Murmann et al. [MGAD19] to train ControlNet [ZRA23] on single view supervised relighting. The network accepts an image (along with its estimated depth map) and a target light direction as input and produces a relit version of the same scene under the desired target lighting.

3. Method

Our method is composed of three main parts. First, we create a 2D relighting neural network with direct control of lighting direction (Sec. 3.1). Second, we use this network to augment a multi-view capture with single lighting into a multi-lighting dataset, by using our relighting network. The resulting dataset can be used to create a radiance field representation of the 3D scene (Sec. 3.2). Finally, we create a relightable radiance field that accounts for inaccuracies in the synthesized relit input images and provides a multi-view consistent lighting solution (Sec. 3.3).

3.1. Single-View Relighting with 2D Diffusion Priors

Relighting a scene captured under a single lighting condition is severely underconstrained, given the lighting/material ambiguity, and thus requires priors about how appearance changes with illumination. Arguably, large DMs must internally encode such priors since they can generate realistic complex lighting effects, but existing architectures do not allow for explicit control over lighting.

We propose to provide explicit control over lighting by fine-tuning a pre-trained Stable Diffusion (SD) [RBL*22] model using ControlNet [ZRA23] on a multi-illumination dataset. As illustrated in Fig. 2, the ControlNet accepts as input an image as well as a target light direction, and produces a relit version of the same scene under the desired lighting. To train the ControlNet, we leverage the dataset of Murmann et al. [MGAD19], which contains $N = 1015$ real indoor scenes captured from a single viewpoint, each lit under $M = 25$ different, controlled lighting directions. We only keep the 18 non-front facing light directions.

3.1.1. Lighting Direction

To capture the scenes using similar light directions, Murmann et al. relied on a camera-mounted directional flash controlled by a servo motor. A pair of diffuse and metallic spheres are also visible in each scene; we leverage the former to obtain the effective lighting directions. Using as target the average of all diffuse spheres produced by the same flash direction, we find the lighting direction $l \in \mathbb{R}^3$

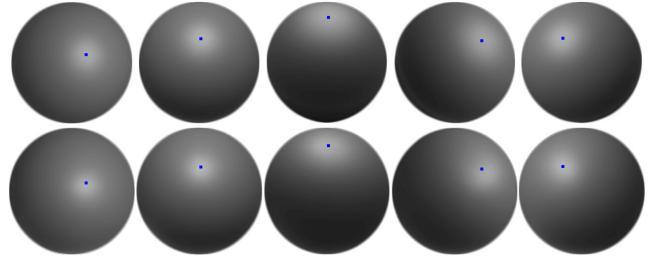


Figure 3: Top row: five diffuse sphere rendered by our optimized lighting direction and shading parameters — the direction is indicated by a blue dot at the point of maximum specular intensity; Bottom row: the corresponding target gray spheres obtained by averaging the diffuse spheres captured in all spheres. We found the lighting directions by minimizing the L_1 distance between the top and bottom row.

which best reproduces this target when rendering a gray ball with a simplistic Phong shading model. More specifically, we minimize the L_1 error when jointly optimizing for an ambient light term and shading parameters (albedo, specular intensity and hardness, as well as a Fresnel coefficient). Fig. 3 illustrates this process.

3.1.2. Controlling Relighting Diffusion

We train ControlNet to predict relit versions of the input image by conditioning it on a target lighting direction. Let us denote a set \mathcal{X} of images of a given scene in the multi-light dataset of Murmann et al. [MGAD19], where each image $\mathbf{X}_k \in \mathcal{X}$ has associated light direction l_k . Our approach, illustrated in Fig. 2, trains on pairs of lighting directions of the same scene (including the identity pair). The denoising objective becomes

$$\mathcal{L}_{2D} = \mathbb{E}_{\epsilon, \mathbf{X}_{t,i}, j} \left[\| g_{\psi}(\mathbf{X}_{t,i}; t, \mathbf{X}_j, \mathbf{D}_j, l_i) - y_t \|_2^2 \right], \quad (1)$$

where $\mathbf{X}_{t,i}$ is the noisy image at timestep $t \in [1, T]$, where $i, j \in [1, M]$, and where ψ are the ControlNet optimizable parameters only. \mathbf{X}_j is another image from the set and \mathbf{D}_j is its depth map (obtained

with the approach of Ke et al. [KOH^{*24}])—both are given as input to the ControlNet subnetwork. In short, the network is trained to denoise input image \mathbf{X}_i given its light direction l_i while conditioned on the image \mathbf{X}_j corresponding to another lighting direction l_j of the same scene. Here, we do not use text conditioning: the empty text string is provided to the network.

Specifically, the light direction l_i is encoded using the first 4 bands of spherical harmonics, following the method of Müller et al. [MESK22]. The resulting vector is added to the timestep embedding prior to feeding it to the layers of ControlNet’s trainable copy.

3.1.3. Improving the Diffusion Quality

Since ControlNet was not specifically designed for relighting, adapting it naively as described above leads to inaccurate colors and a loss in contrast (see Fig. 4), as well as distorted edges (see Fig. 5). These errors also degrade multi-view consistency.



Figure 4: Importance of post-relighting color and contrast adjustments. Left: input image. Middle: naive ControlNet relighting; the bottle has the wrong color and the contrast is poor. Right: our relighting after training with [LLY23] and after color-matching the input.

We adopt two strategies to improve coloration and contrast. First, we follow the recommendations of [LLY23] to improve image brightness—we found them to also help for color. In particular, using the “v-parameterized” objective $y_t = \sqrt{\bar{\alpha}_t} \cdot \epsilon - \sqrt{1 - \bar{\alpha}_t} \cdot x$, instead of the more usual $y_t = \epsilon$, proved critical; in this equation, $1 - \bar{\alpha}_t$ gives the variance of the noise at timestep t . Second, after sampling, we color-match predictions to the input image to compensate for the difference between the color distribution of the training data and that of the scene. This is done by subtracting the per-channel mean and dividing by the standard deviation for the prediction, then adding the mean and standard deviation of the input, in the LAB colorspace. This is computed over all 18 lighting conditions together (i.e., the mean over all lighting directions) to conserve relative brightness across all conditions. Fig. 4 shows the effect of these changes; without them, the bottle is blue instead of green and overall contrast is poor.

To correct the distorted edges, we adapt the asymmetric autoencoder approach of Zhu et al. [ZFC^{*23}], which consists in conditioning the latent space decoder with the (masked) input image for the inpainting task. In our case, we ignore the masking and fine-tune the decoder on the multi-illumination dataset [MGAD19]. At each fine-tuning step, we encode an image and condition the decoder on an image from the same scene with another random lighting direction. The decoder is fine-tuned with the Adam optimizer at



Figure 5: Importance of conserving edge sharpness when relighting. Left: input image. Middle: naive ControlNet relighting; note how the edges do not match the input and how the text is illegible. Right: our final relighting after fine-tuning the conditional decoder network from [ZFC^{*23}].

a learning rate of 10^{-4} for 20k steps when training at resolution 768×512 and 50k steps at resolution 1536×1024 . Note that this step is independent of the ControlNet training. Fig. 5 shows the effect of these changes; note how the edges are wobbly and the text is illegible without them.

Example relighting results obtained using our 2D relighting network on images outside of the dataset are shown in Fig. 6. Observe how the relit images produced by our method are highly realistic and light directions are consistently reproduced across scenes. A naive solution for radiance field relighting would be to apply this 2D network to each synthesized novel view. However, the ControlNet is not multi-view consistent, and such a naive solution results in significant flickering. Please see the accompanying video for a clear illustration.

3.2. Augmenting Multi-View/Single-Lighting Datasets

Given a multi-view set \mathcal{I} of images of a scene captured under the same lighting (suitable for training a radiance field model), we now leverage our light-conditioned ControlNet model to synthetically relight each image in \mathcal{I} . We assume the 3D pose of each image $\mathbf{I}_i \in \mathcal{I}$ is known a priori, for example via Colmap [SF16, SZPF16]. We then simply relight each $\mathbf{I}_i \in \mathcal{I}$ to the corresponding 18 known light directions in the dataset from Murmann et al. [MGAD19] (excluding the directions where the flash points forward), (see Sec. 3.1). We now have a full multi-lighting, multi-view dataset. This process is illustrated in Fig. 7.

3.3. Training a Lighting-Consistent Radiance Field

Given the generated multi-light, multi-view dataset, we now describe our solution to provide a relightable radiance field. In particular, we build on the 3DGS framework of Kerbl et al. [KKLD23]. Our requirements are twofold: first, define an augmented radiance field that can represent lighting conditions from different lighting directions; second, allow direct control of the lighting direction used for relighting.

The original 3DGS [KKLD23] radiance field uses spherical harmonics (SH) to represent view-dependent illumination. To encode varying illumination, we replace the SH coefficients with a 3-layer MLP c_θ of width 128 which takes as input the light direction along

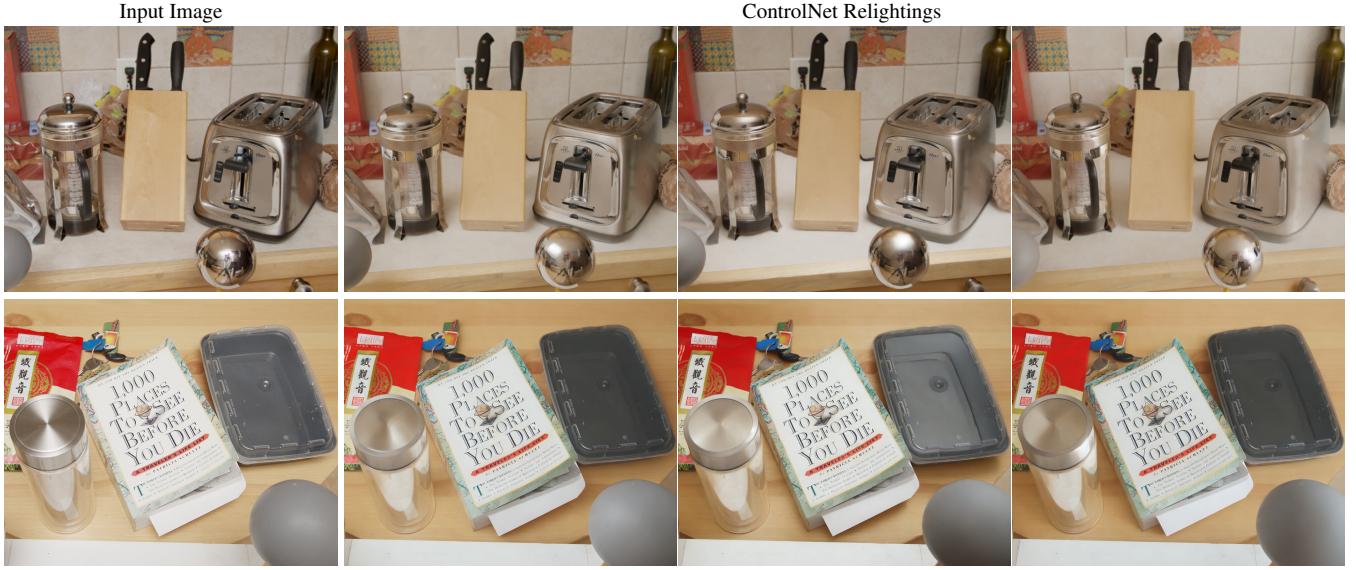


Figure 6: Relighting results with our light-conditioned ControlNet. From a single input image (left column), the ControlNet can generate realistic relit versions for different target light directions (other columns). Please notice realistic changes in highlights for different light directions (top row), as well as the synthesis of cast shadows (bottom row).

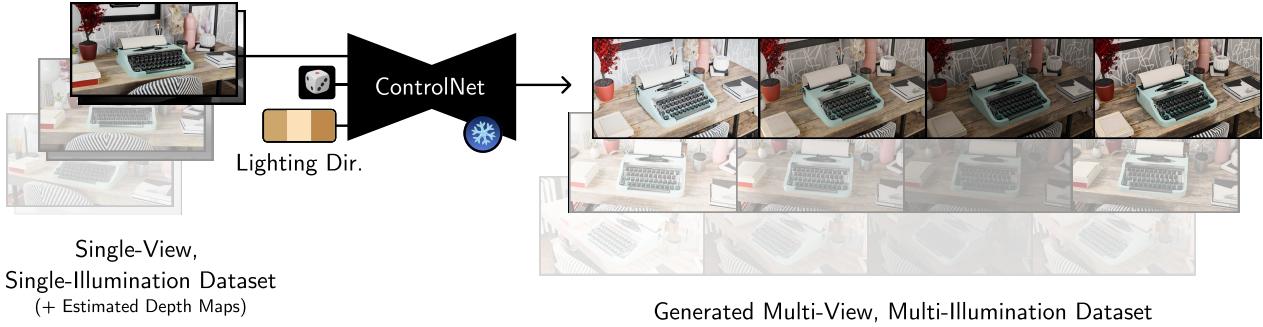


Figure 7: Given a multi-view, single-illumination dataset we use our relighting ControlNet to generate a multi-view, multi-illumination dataset.

with the viewing direction. Both vectors have a size of 16 after encoding.

Since light directions are computed with respect to a local camera reference frame (c.f. Sec. 3.1), we subsequently register them to the world coordinate system (obtained from Colmap) by rotating them according to their (known) camera rotation parameters:

$$l' = R_i l, \quad (2)$$

where R_i is the 3×3 camera-to-world rotation matrix of image \mathbf{I}_i from its known pose.

We condition the MLP with the spherical harmonics encoding of the globally consistent lighting direction l' , which enables training a 3DGS representation on our multi-lighting dataset. While this strategy works well for static images, it results in inconsistent lighting across views despite accounting for camera rotation in Eq. 2. Radi-

ance fields like 3DGS rely on multi-view consistency, and breaking it introduces additional floaters and holes in surfaces.

To allow the neural network to account for this inconsistency and correct accordingly, we optimize a per-image auxiliary latent vector a of size 128. Similar approaches for variable appearance have been used for NeRFs [MBRS*21]. Therefore, in addition to the lighting direction l' , we condition the MLP with per-view auxiliary parameters a :

$$c(o, d) = \sum_{g=1}^G w_g c_\theta(x_g, d | a_v, l'), \quad (3)$$

where $g \in [1, G]$ sums over the G gaussians (see [KKLD23]), x_g/w_g are their features/weights, d is the view direction, o the ray origin, and c is the predicted pixel color. Note that for novel views at

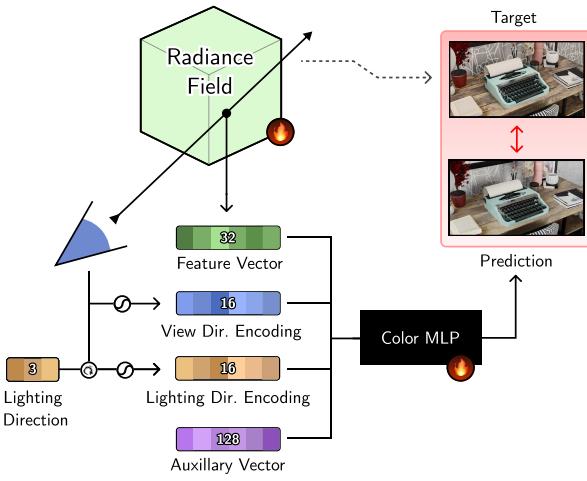


Figure 8: Overview of our radiance field training scheme. To alleviate potential inconsistencies in lighting directions, we condition our 3DGS-based radiance field both on the illumination direction encoding and on optimized auxiliary vectors (one per training image). These vectors model the differences between predictions and let us fit each view to convergence.

inference, we use as latent vector the mean of all training view latents i.e. $\mathbb{E}_v[a_v]$.

We first train 3DGS with the unlit images as a “warmup” stage for 5K iterations, then train the full multi-illumination solution for another 25K iterations, using all 18 back-facing light directions (see Sec. 3.1). The multi-illumination nature of the training results in an increase in “floaters”. As observed by Philip and Deschaindre [PD23], floaters are often present close to the input cameras; the explicit nature of 3DGS allows us to reduce these effectively. In particular, we calculate a z_{near} value for all cameras by taking the z value of the 1st percentile of nearest SfM points and scaling this value down by 0.9. During training, at each step, all gaussian primitives that project within the view frustum of a camera but are located in front of its z_{near} plane are culled. Finally, given the complexity of modeling variable lighting, we observed that the optimization sometimes converges to blurry results. To counter this, we overweight three front-facing views (left, right, and center), by optimizing for one of these views every three iterations. This provides marginal improvement in results; all images shown are computed with this method, but it is optional.

The full method for relightable radiance fields is shown in Fig. 8. At inference, we can directly choose a lighting direction, and use efficient 3DGS rendering for interactive updates with modified lighting. Our latent vectors and floater removal remove most, but not all, artifacts introduced by the multi-view inconsistencies; this can be seen in the ablations at the end of the supplemental video.

4. Results and Evaluation

Our method was implemented by leveraging publicly available implementations of ControlNet [ZRA23] and 3DGS [KKLD23]. We

use Stable Diffusion [RBL^{*}22] v2.1 as a backbone. Our source code and datasets will be released upon publication.

We first present the results of our 3D relightable radiance field, both for synthetic and real-world scenes. We then present a quantitative and qualitative evaluation of our method by comparing it to previous work and finally present an ablation of the auxiliary vector a from Sec. 3.3.

4.1. Test Datasets

Since there are no real *multi-view* multi-illumination indoor datasets of full scenes available for our evaluation, we use synthetic scenes to allow quantitative evaluation. For this purpose, we designed 4 synthetic test scenes (KITCHEN, LIVINGROOM, OFFICE, BEDROOM). They were created in Blender by downloading artist-made 3D rooms from Evermotion and modifying them to increase clutter: in each room, we gathered objects and placed them on a table or a countertop. We also created simpler, diffuse-only versions to evaluate how scene clutter affects the relighting results. For each synthetic scene, we first built a standard multi-view (single-lighting) dataset consisting of 4 camera sweeps (left-to-right, at varying elevations) of 50 frames for training and one (at a different elevation) of 100 frames for testing. We simulated the light direction of the 2D training dataset with a spotlight with intensity of 2 kW and radius 0.1 located on top of the camera and pointing away from it. We used the same set of camera flash directions as in the dataset of Murmann et al. [MGAD19]. We then render all frames in 736×512 using the Cycles path tracer. Please note that the effective lighting direction will be dependent on the exact configuration of the room. This configuration is our best effort to produce a ground truth usable for comparison.

In addition, we also captured a set of real scenes (KETTLE, HOT PLATES, PAINT GUN, CHEST OF DRAWERS and GARAGE WALL), for which we performed a standard radiance-field multi-view capture, by taking between 90–150 images of the environment, in an approximate sphere (or hemisphere) around the scene center of interest.

4.2. 3D Relighting Results

We begin by showing qualitative results on the set of real scenes that we captured. Here, we used a resolution of 1536×1024 , training for 150K iterations. We show qualitative results for these scenes using our 3D relightable radiance field in Fig. 9. In addition, we also show results for two scenes from the MipNeRF360 dataset, namely COUNTER and ROOM.

As our method is lightweight and only adds a small MLP over the core 3DGS architecture, it runs interactively for both novel view synthesis and relighting at 30fps on an A6000 GPU. Memory usage is comparable to the original 3DGS. Please see the video for interactive relighting results on these scenes and additional synthetic scenes. We see that our method produces realistic and plausible relighting results. Also, note that our solution is temporally consistent.

4.3. Evaluation

Baselines. We compare our results to the method of [PMGD21] which is specifically designed for complete scenes, Ten-

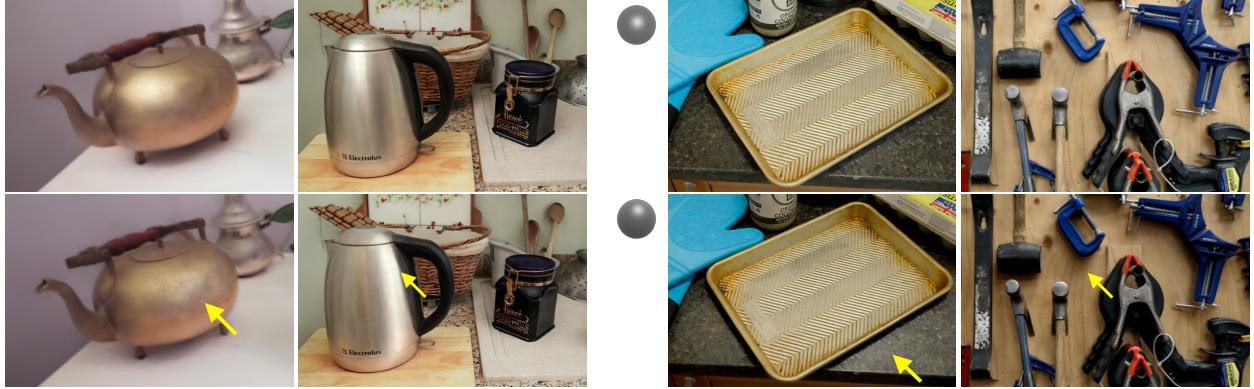


Figure 9: Qualitative relighting results for the real scenes, from left to right: CHEST OF DRAWERS, KETTLE, MIPNERF ROOM and GARAGE WALL, for a moving light source. The lighting direction is indicated in the gray ball in the lower right. Please see the supplemental video for more results. Please note how the highlights (left group) and shadows (right group) have changed.



Figure 10: Qualitative comparison on real scene KETTLE. From left to right, from the same viewpoint: input lighting condition (view reconstructed using 3D Gaussian Splatting), target lighting, our relighting, Philip et al. [PMGD21] relighting. Top and bottom rows are two different lighting conditions. Philip et al. [PMGD21] exhibits much more geometry and shading artifacts compared to our method; in particular imprecise MVS preprocessing results in missing geometry.

Method →	Ours			Outcast [GRP22]			R3DGS [GGL*23]			TensoIR [JLX*23]		
Scene ↓ / Metrics	PSNR↑	LPIPS↓	SSIM↑	PSNR↑	LPIPS↓	SSIM↑	PSNR↑	LPIPS↓	SSIM↑	PSNR↑	LPIPS↓	SSIM↑
Simple Bedroom	20.57	0.156	0.868	17.24	0.207	0.808	17.79	0.174	0.830	15.77	0.471	0.595
Simple Kitchen	17.45	0.154	0.855	17.91	0.205	0.822	18.55	0.197	0.807	20.52	0.382	0.701
Simple Livingroom	22.12	0.136	0.884	21.09	0.125	0.878	20.34	0.166	0.857	17.45	0.444	0.598
Simple Office	18.59	0.131	0.868	18.97	0.196	0.811	20.40	0.173	0.808	18.22	0.446	0.644
Complex Bedroom	17.70	0.145	0.791	15.26	0.221	0.694	16.69	0.186	0.741	14.42	0.434	0.555
Complex Kitchen	19.28	0.152	0.811	18.44	0.178	0.771	19.28	0.168	0.755	16.70	0.471	0.533
Complex Livingroom	18.61	0.163	0.800	17.94	0.187	0.783	18.39	0.175	0.770	16.82	0.382	0.602
Complex Office	20.20	0.096	0.858	17.22	0.169	0.781	18.93	0.144	0.776	15.78	0.468	0.529

Table 1: Quantitative results of our 3D relighting on the synthetic datasets (where ground truth is available), compared to previous work, from left to right: OutCast [GRP22] (run on individual images from 3DGS [KKLD23]), Relightable3DGaussians [GGL*23], and TensoIR [JLX*23]. Arrows indicate higher/lower (\uparrow / \downarrow) is better. Results are color coded by best, second- and third-best.



Figure 11: We show comparative results of our method of synthetic scenes where the (approximate) ground truth is available (left), and compare to previous methods. Our approach is closer to the ground truth lighting, capturing the overall appearance in a realistic manner.

soIR [JLX^{*}23] and RelightableGaussians [GGL^{*}23]. Given that most other methods do not handle full scenes well, we also create a new baseline, by first training 3DGS [KLKD23] on the input data and render a test path using novel view synthesis; We then use Out-Cast [GRP22] to relight each individual rendered frame using the target direction. We trained TensoIR [JLX^{*}23] using the default configuration but modified the “density_shift” parameter from -10 to -8 to achieve best results on our data. For Relightable 3D Gaussians [GGL^{*}23], we train their “Stage 1” for 30K iterations and “Stage 2” for an additional 10K to recover the BRDF parameters. We then relight the scenes using 360° environment maps rendered in Blender using a generic empty room and a similar camera/flash setup for ground truth. Finally, to improve the baselines we normalize the predictions of all methods; we first subtract the channel-wise mean and divide out the channel-wise standard deviation, and then multiply and add the corresponding parameters of the ground truths. These operations are performed in LAB space for all methods.

Experimental methodology. We use our synthetic test scenes for providing quantitative results. To compare our method, we rendered 200 novel views with 18 different lighting directions to evaluate the relighting quality for each method by computing standard image quality metrics. Given the complexity of setup for [PMGD21], we only show qualitative results for 1 real scene in Fig. 10. Here, our method was trained at 768×512 resolution for 200k iterations, with a batch size of 8 and a learning rate of 10^{-4} .

Results. We present quantitative results in Table 1. We present per-scene results on the following image quality metrics: PSNR, SSIM, and LPIPS [ZIE^{*}18]. The results demonstrate that our method outperforms all others in all but a few scenarios, where it still achieves competitive performance.

Qualitative comparisons are shown in Fig. 11; on the left we show the ground truth relit image rendered in Blender, and we then show our results, as well as those from Outcast [GRP22], Relightable 3D Gaussians [GGL^{*}23] and TensoIR [JLX^{*}23]. Please refer to the supplementary HTML viewer for more results. We clearly see that our method is closer to the ground truth, visually confirming the quantitative results in Tab. 1. TensoIR has difficulty reconstructing the geometry, and Relightable 3D Gaussians tend to have a “splotchy” look due to inaccurate normals. Outcast has difficulty with the overall lighting condition and can add incorrect shadows, but in many cases produces convincing results since it operates in image space. Our results show that by using the diffusion prior we manage to achieve realistic relighting, surpassing the state of the art.

Our method was trained for indoor scenes; Fig. 13 gives additional ControlNet results on out-of-distribution samples, showing that it can generalize to some extent to unseen scenes and lighting conditions, although the realism is lower than for in-distribution samples.

5. Conclusion

We have presented the first method to effectively leverage the strong prior of large generative diffusion models in the context of radiance field relighting. Rather than relying on accurate geometry, material and/or lighting estimation, our approach models realistic illumination directly, by leveraging a general-purpose single-view,

multi-illumination dataset and fine-tuning a large pretrained generative model. Our results show that we can synthesize realistic relighting of captured scenes, while allowing interactive novel-view synthesis by building on such priors. Our method shows levels of realism for relighting that surpass the state of the art for cluttered indoor scenes (as opposed to isolated objects).



Figure 12: Example limitations of our approach, with our prediction (top) vs ground truth (bottom). Our ControlNet mistakenly produces a shadow at the top of the image while there should not be any (red arrow), presumably assuming the presence of another top shelf. Additionally, the highlight position is somewhat incorrect (yellow arrow), ostensibly because we define light direction in a manner that is not fully physically accurate.

One limitation of the proposed method is that it does not enforce physical accuracy: the target light direction is noisy and the ControlNet relies mostly on its powerful Stable Diffusion prior to relight rather than performing physics-based reasoning. For example, Fig. 12 shows that ControlNet can hallucinate shadows due to unseen geometry, while there should not be any. Given that we define light direction in a manner that is not fully physically accurate, the positioning of highlight can be inaccurate, as is also shown in Fig. 12. In addition, the appearance embeddings can correct for global inconsistencies indirectly and do not explicitly rely on the learned 3D representation of the radiance field. Our method does not always remove or move shadows in a fully accurate physically-based manner. While our method clearly demonstrates that 2D diffusion model priors can be used for realistic relighting, the ability to perform more complex relighting—rather than just changing light direction—requires significant future research, e.g., by using more general training data as well as ways to encode and decode complex lighting.

An interesting direction for future work would be trying to enforce multi-view consistency more explicitly in ControlNet, e.g. by leveraging single-illumination multi-view data. Another interesting direction is to develop solutions that would guide the predicted relighting making it more accurate, leveraging the 3D geometric information available in the radiance field more explicitly.

Acknowledgements This research was funded by the ERC Advanced grant FUNGRAPH No 788065 <http://fungraph>.

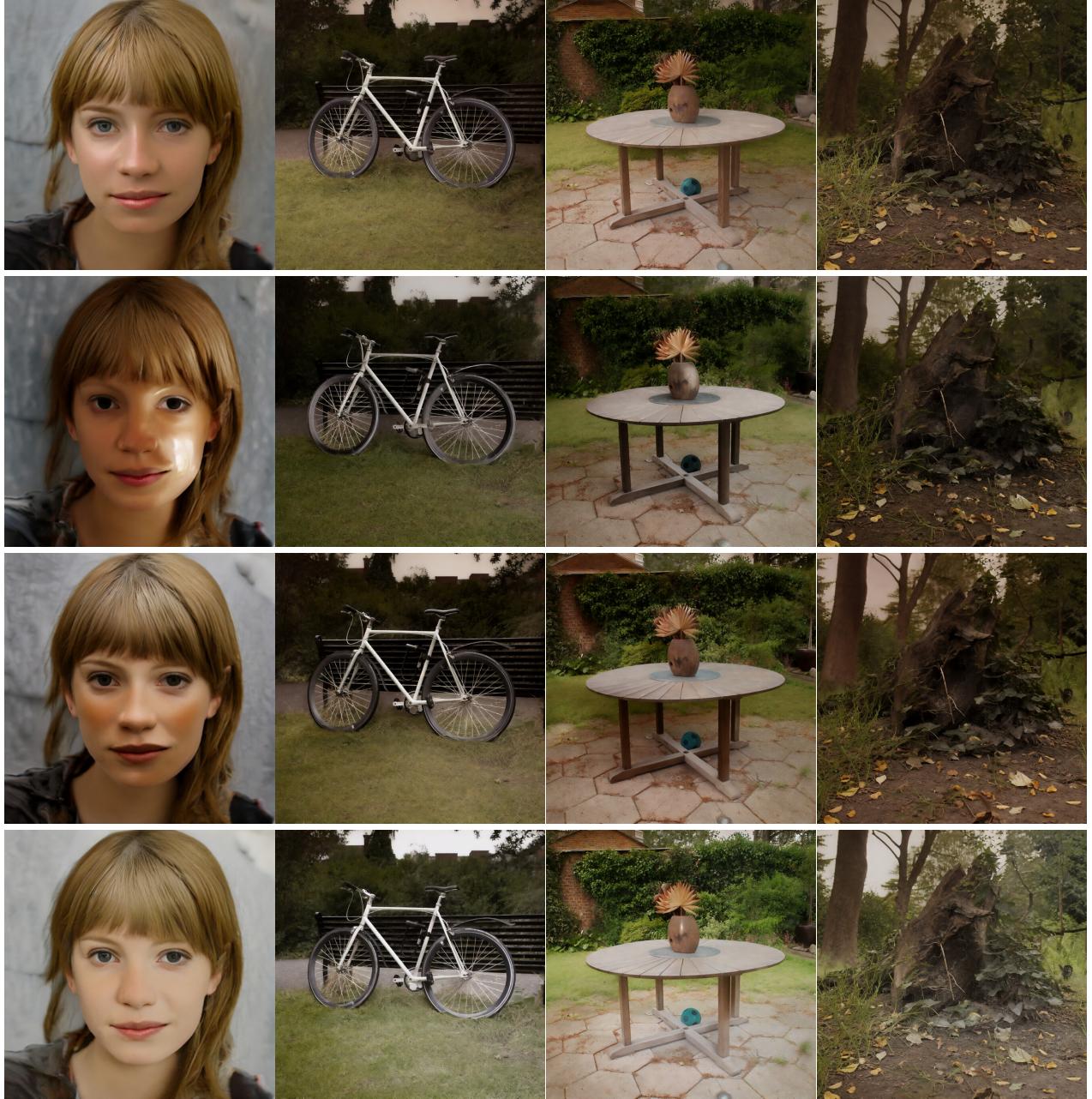


Figure 13: We show the results of our 2D relighting network on out-of-distribution images (StyleGAN-generated woman and MipNeRF360 BICYCLE, GARDEN, and STUMP). On human faces, ControlNet may change the expression as well as the lighting, or create excessive shininess; on outdoor scenes, while the overall lighting direction is plausible, the network fails to generate sufficiently hard shadows.

inria.fr/, supported by NSERC grant DGPIN 2020-04799 and the Digital Research Alliance Canada. The authors are grateful to Adobe and NVIDIA for generous donations, and the OPAL infrastructure from Université Côte d’Azur. Thanks to Georgios Kopanas and Frédéric Fortier-Chouinard for helpful advice.

References

- [BBJ^{*}21] BOSS M., BRAUN R., JAMPANI V., BARRON J. T., LIU C., LENSCHE H.: Nerd: Neural reflectance decomposition from image collections. In *IEEE/CVF Int. Conf. Comput. Vis.* (2021). [2](#), [3](#)
- [BEK^{*}22] BOSS M., ENGELHARDT A., KAR A., LI Y., SUN D., BARRON J. T., LENSCHE H. P., JAMPANI V.: SAMURAI: Shape and material from unconstrained real-world arbitrary image collections. In *Adv. Neural Inform. Process. Syst.* (2022). [3](#)

- [BF24] BHATTAD A., FORSYTH D. A.: Stylitgan: Prompting stylegan to produce new illumination conditions. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog.* (2024). 3
- [BJB*21] BOSS M., JAMPANI V., BRAUN R., LIU C., BARRON J., LENSCHE H.: Neural-pil: Neural pre-integrated lighting for reflectance decomposition. In *Adv. Neural Inform. Process. Syst.* (2021). 2, 3
- [BMHF23] BHATTAD A., MCKEE D., HOIEM D., FORSYTH D.: Stylegan knows normal, depth, albedo, and more. 3
- [BMT*21] BARRON J. T., MILDENHALL B., TANCIK M., HEDMAN P., MARTIN-BRULLA R., SRINIVASAN P. P.: Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In *IEEE/CVF Int. Conf. Comput. Vis.* (2021). 2
- [BMV*22] BARRON J. T., MILDENHALL B., VERBIN D., SRINIVASAN P. P., HEDMAN P.: Mip-nerf 360: Unbounded anti-aliased neural radiance fields. *IEEE/CVF Conf. Comput. Vis. Pattern Recog.* (2022). 2
- [CKK23] CHOI C., KIM J., KIM Y. M.: IBL-NeRF: Image-based lighting formulation of neural radiance fields. *Comput. Graph. Forum* 42, 7 (2023). 3
- [CXG*22] CHEN A., XU Z., GEIGER A., YU J., SU H.: Tensorf: Tensorial radiance fields. In *Eur. Conf. Comput. Vis.* (2022). 2
- [DHT*00] DEBEVEC P., HAWKINS T., TCHOU C., DUKER H.-P., SAROKIN W., SAGAR M.: Acquiring the reflectance field of a human face. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques* (2000), pp. 145–156. 1
- [FRV*23] FUTSCHIK D., RITLAND K., VECORE J., FANELLO S., ORTS-ESCOLANO S., CURLESS B., SÝKORA D., PANDEY R.: Controllable light diffusion for portraits. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog.* (2023). 2
- [GAA*23] GAL R., ARAR M., ATZMON Y., BERMANO A. H., CHECHIK G., COHEN-OR D.: Encoder-based domain tuning for fast personalization of text-to-image models. *ACM Trans. Graph.* 42, 4 (2023), 1–13. 3
- [GCD*20] GAO D., CHEN G., DONG Y., PEERS P., XU K., TONG X.: Deferred neural lighting: free-viewpoint relighting from unstructured photographs. *ACM Trans. Graph.* 39, 6 (nov 2020). 3
- [GGI*23] GAO J., GU C., LIN Y., ZHU H., CAO X., ZHANG L., YAO Y.: Relightable 3d gaussian: Real-time point cloud relighting with brdf decomposition and ray tracing. *arXiv:2311.16043* (2023). 2, 3, 8, 10
- [GRP22] GRIFFITHS D., RITSCHEL T., PHILIP J.: Outcast: Outdoor single-image relighting with cast shadows. In *Comput. Graph. Forum* (2022), vol. 41, pp. 179–193. 3, 8, 10
- [HHM22] HASSELGREN J., HOFMANN N., MUNKBERG J.: Shape, light, and material decomposition from images using monte carlo rendering and denoising. *Adv. Neural Inform. Process. Syst.* 35 (2022), 22856–22869. 3
- [HJA20] HO J., JAIN A., ABBEEL P.: Denoising diffusion probabilistic models. In *Adv. Neural Inform. Process. Syst.* (2020). 3
- [JLX*23] JIN H., LIU I., XU P., ZHANG X., HAN S., BI S., ZHOU X., XU Z., SU H.: Tensori: Tensorial inverse rendering. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog.* (2023). 2, 3, 8, 10
- [JTL*24] JIANG Y., TU J., LIU Y., GAO X., LONG X., WANG W., MA Y.: Gaussianshader: 3d gaussian splatting with shading functions for reflective surfaces, 2024. 3
- [KE19] KANAMORI Y., ENDO Y.: Relighting humans: occlusion-aware inverse rendering for full-body human images. *ACM Trans. Graph.* 37, 6 (2019). 2
- [KKLD23] KERBL B., KOPANAS G., LEIMKÜHLER T., DRETTAKIS G.: 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.* 42, 4 (July 2023). 2, 5, 6, 7, 8, 10
- [KOH*24] KE B., OBUKHOV A., HUANG S., METZGER N., DAUDT R. C., SCHINDLER K.: Repurposing diffusion-based image generators for monocular depth estimation. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog.* (2024). 5
- [LCL*23] LIANG R., CHEN H., LI C., CHEN F., PANNEER S., VIJAYKUMAR N.: Envidr: Implicit differentiable renderer with neural environment lighting. In *IEEE/CVF Int. Conf. Comput. Vis.* (2023). 3
- [LGF*22] LI Q., GUO J., FEI Y., LI F., GUO Y.: Neulighting: Neural lighting for free viewpoint outdoor scene relighting with unconstrained photo collections. In *SIGGRAPH Asia 2022 Conference Papers* (2022). 3
- [LGZ*20] LIU A., GINOSAR S., ZHOU T., EFROS A. A., SNAVELY N.: Learning to factorize and relight a city. In *Eur. Conf. Comput. Vis.* (2020). 2
- [LLLY23] LIN S., LIU B., LI J., YANG X.: Common diffusion noise schedules and sample steps are flawed. *arXiv preprint arXiv:2305.08891* (2023). 5
- [LLZ*20] LIU D., LONG C., ZHANG H., YU H., DONG X., XIAO C.: Arshadowgan: Shadow generative adversarial network for augmented reality in single light scenes. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog.* (2020). 2
- [LSB*22] LI Z., SHI J., BI S., ZHU R., SUNKAVALLI K., HAŠAN M., XU Z., RAMAMOORTHI R., CHANDRAKAR M.: Physically-based editing of indoor scene lighting from a single image. In *Eur. Conf. Comput. Vis.* (2022). 2
- [LSY*21] LAGUNAS M., SUN X., YANG J., VILLEGRAS R., ZHANG J., SHU Z., MASIA B., GUTIERREZ D.: Single-image full-body human relighting. *Eur. Graph. Symp. Render.* (2021). 2
- [LWC*23] LI Z., WANG L., CHENG M., PAN C., YANG J.: Multi-view inverse rendering for large-scale real-world indoor scenes. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog.* (2023). 3
- [LWL*23] LIU Y., WANG P., LIN C., LONG X., WANG J., LIU L., KOMURA T., WANG W.: Nero: Neural geometry and brdf reconstruction of reflective objects from multiview images. *ACM Trans. Graph.* (2023). 3
- [LZF*24] LIANG Z., ZHANG Q., FENG Y., SHAN Y., JIA K.: Gs-ir: 3d gaussian splatting for inverse rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2024). 3
- [MBRS*21] MARTIN-BRULLA R., RADWAN N., SAJJADI M. S. M., BARRON J. T., DOSOVITSKIY A., DUCKWORTH D.: NeRF in the Wild: Neural Radiance Fields for Unconstrained Photo Collections. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog.* (2021). 6
- [MESK22] MÜLLER T., EVANS A., SCHIED C., KELLER A.: Instant neural graphics primitives with a multiresolution hash encoding. *ACM Trans. Graph.* 41, 4 (July 2022), 102:1–102:15. 2, 5
- [MGAD19] MURMANN L., GHARBI M., AITTALA M., DURAND F.: A multi-illumination dataset of indoor object appearance. In *IEEE/CVF Int. Conf. Comput. Vis.* (2019). 2, 4, 5, 7
- [MHS*22] MUNKBERG J., HASSELGREN J., SHEN T., GAO J., CHEN W., EVANS A., MÜLLER T., FIDLER S.: Extracting Triangular 3D Models, Materials, and Lighting From Images. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog.* (June 2022), pp. 8280–8290. 3
- [MST*20] MILDENHALL B., SRINIVASAN P. P., TANCIK M., BARRON J. T., RAMAMOORTHI R., NG R.: Nerf: Representing scenes as neural radiance fields for view synthesis. In *Eur. Conf. Comput. Vis.* (2020). 1, 2
- [NDDJK21] NIMIER-DAVID M., DONG Z., JAKOB W., KAPLANYAN A.: Material and lighting reconstruction for complex indoor scenes with texture-space differentiable rendering. *Comput. Graph. Forum* (2021). 3
- [PD23] PHILIP J., DESCHAINTRE V.: Floaters no more: Radiance field gradient scaling for improved near-camera training. In *EGSR Conference proceedings DL-track* (2023), The Eurographics Association. 7
- [PEL*23] PODELL D., ENGLISH Z., LACEY K., BLATTMANN A., DOCKHORN T., MÜLLER J., PENNA J., ROMBACH R.: Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952* (2023). 3
- [PGZ*19] PHILIP J., GHARBI M., ZHOU T., EFROS A. A., DRETTAKIS G.: Multi-view relighting using a geometry-aware network. *ACM Trans. Graph.* 38, 4 (2019), 78–1. 2, 3

- [PLMZ23] PAPANTONIOU F. P., LATTAS A., MOSCHOGLOU S., ZAFEIRIOU S.: Relightify: Relightable 3d faces from a single image via diffusion models. In *IEEE/CVF Int. Conf. Comput. Vis.* (2023). 2
- [PMGD21] PHILIP J., MORGENTHALER S., GHARBI M., DRETTAKIS G.: Free-viewpoint indoor neural relighting from multi-view stereo. *ACM Trans. Graph.* 40, 5 (2021), 1–18. 2, 3, 7, 8, 10
- [PTS23] PONGLERTNAPAKORN P., TRITRONG N., SUWAJANAKORN S.: Difareli: Diffusion face relighting. In *IEEE/CVF Int. Conf. Comput. Vis.* (2023). 2
- [RBL*22] ROMBACH R., BLATTMANN A., LORENZ D., ESSER P., OMMER B.: High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog.* (2022). 2, 3, 4, 7
- [RES*22] RUDNEV V., ELGHARIB M., SMITH W., LIU L., GOLYANIK V., THEOBALT C.: Nerf for outdoor scene relighting. In *Eur. Conf. Comput. Vis.* (2022). 3
- [RLJ*23] RUIZ N., LI Y., JAMPANI V., PRITCH Y., RUBINSTEIN M., ABERMAN K.: Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog.* (2023), pp. 22500–22510. 3
- [SDWMG15] SOHL-DICKSTEIN J., WEISS E., MAHESWARANATHAN N., GANGULI S.: Deep unsupervised learning using nonequilibrium thermodynamics. 3
- [SDZ*21] SRINIVASAN P. P., DENG B., ZHANG X., TANCIK M., MILDENHALL B., BARRON J. T.: Nerv: Neural reflectance and visibility fields for relighting and view synthesis. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog.* (2021). 3
- [SF16] SCHÖNBERGER J. L., FRAHM J.-M.: Structure-from-motion revisited. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog.* (2016). 2, 5
- [S JL*23] SHARMA P., JAMPANI V., LI Y., JIA X., LAGUN D., DURAND F., FREEMAN W. T., MATTHEWS M.: Alchemist: Parametric control of material properties with diffusion models. *arXiv preprint arXiv:2312.02970* (2023). 3
- [SKCJ18] SENGUPTA S., KANAZAWA A., CASTILLO C. D., JACOBS D. W.: Sfsnet: Learning shape, reflectance and illuminance of faces in the wild. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog.* (2018). 2
- [SLZ*22] SHENG Y., LIU Y., ZHANG J., YIN W., OZTIRELI A. C., ZHANG H., LIN Z., SHECHTMAN E., BENES B.: Controllable shadow generation using pixel height maps. In *Eur. Conf. Comput. Vis.* (2022). 2
- [SME20] SONG J., MENG C., ERMON S.: Denoising diffusion implicit models. In *Int. Conf. Learn. Represent.* (2020). 3
- [SWW*23] SHI Y., WU Y., WU C., LIU X., ZHAO C., FENG H., LIU J., ZHANG L., ZHANG J., ZHOU B., DING E., WANG J.: Gir: 3d gaussian inverse rendering for relightable scene factorization. *Arxiv* (2023). 3
- [SYH*17] SHU Z., YUMER E., HADAP S., SUNKAVALLI K., SHECHTMAN E., SAMARAS D.: Neural face editing with intrinsic image disentangling. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog.* (2017). 2
- [SZB21] SHENG Y., ZHANG J., BENES B.: Ssn: Soft shadow network for image compositing. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog.* (2021). 2
- [SZPF16] SCHÖNBERGER J. L., ZHENG E., POLLEFEYS M., FRAHM J.-M.: Pixelwise view selection for unstructured multi-view stereo. In *Eur. Conf. Comput. Vis.* (2016). 5
- [TQE*21] TÜRE M., ÇIKLABAKKAL M. E., ERDEM A., ERDEM E., SATILMIŞ P., AKYÜZ A. O.: From noon to sunset: Interactive rendering, relighting, and recolouring of landscape photographs by modifying solar position. In *Comput. Graph. Forum* (2021), vol. 40, pp. 500–515. 2
- [TDMS*23] TOSCHI M., DE MATTEO R., SPEZIALETTI R., DE GREGORIO D., DI STEFANO L., SALTI S.: Relight my nerf: A dataset for novel view synthesis and relighting of real world objects. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog.* (June 2023), pp. 20762–20772. 3
- [TTM*22] TEWARI A., THIES J., MILDENHALL B., SRINIVASAN P., TRETSCHK E., YIFAN W., LASSNER C., SITZMANN V., MARTIN-BRUALLA R., LOMBARDI S., ET AL.: Advances in neural rendering. In *Comput. Graph. Forum* (2022), vol. 41, Wiley Online Library, pp. 703–735. 2
- [Ull79] ULLMAN S.: The interpretation of structure from motion. *Proceedings of the Royal Society of London. Series B, Biological sciences* 203, 1153 (January 1979), 405–426. 2
- [VZG*23] VALENÇA L., ZHANG J., GHARBI M., HOLD-GEOFFROY Y., LALONDE J.-F.: Shadow harmonization for realistic compositing. In *SIGGRAPH Asia 2023 Conference Papers* (2023). 2
- [WSG*23] WANG Z., SHEN T., GAO J., HUANG S., MUNKBERG J., HASSELGREN J., GOJCIC Z., CHEN W., FIDLER S.: Neural fields meet explicit geometric representations for inverse rendering of urban scenes. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog.* (2023). 3
- [WSLG23] WU T., SUN J.-M., LAI Y.-K., GAO L.: De-nerf: Decoupled neural radiance fields for view-consistent appearance editing and high-frequency environmental relighting. In *ACM SIGGRAPH* (2023). 3
- [WZL*08] WANG Y., ZHANG L., LIU Z., HUA G., WEN Z., ZHANG Z., SAMARAS D.: Face relighting from a single image under arbitrary unknown lighting conditions. *IEEE Trans. Pattern Anal. Mach. Intell.* 31, 11 (2008), 1968–1984. 2
- [XZC*23] XU Y., ZOSS G., CHANDRAN P., GROSS M., BRADLEY D., GOTARDO P.: Renerf: Relightable neural radiance fields with nearfield lighting. In *IEEE/CVF Int. Conf. Comput. Vis.* (2023). 3
- [YME*20] YU Y., MEKA A., ELGHARIB M., SEIDEL H.-P., THEOBALT C., SMITH W. A.: Self-supervised outdoor scene relighting. In *Eur. Conf. Comput. Vis.* (2020). 2
- [YS22] YU Y., SMITH W. A. P.: Outdoor inverse rendering from a single image using multiview self-supervision. *IEEE Trans. Pattern Anal. Mach. Intell.* 44, 7 (2022), 3659–3675. 2
- [YZL*22] YAO Y., ZHANG J., LIU J., QU Y., FANG T., MCKINNON D., TSIN Y., QUAN L.: Neilf: Neural incident light field for physically-based material estimation. In *Eur. Conf. Comput. Vis.* (2022). 3
- [ZCD*23] ZENG C., CHEN G., DONG Y., PEERS P., WU H., TONG X.: Relighting neural radiance fields with shadow and highlight hints. In *ACM SIGGRAPH 2023 Conference Proceedings* (2023). 3
- [ZDP*24] ZENG C., DONG Y., PEERS P., KONG Y., WU H., TONG X.: Dilightnet: Fine-grained lighting control for diffusion-based image generation. In *ACM SIGGRAPH 2024 Conference Proceedings* (2024). 3
- [ZFC*23] ZHU Z., FENG X., CHEN D., BAO J., WANG L., CHEN Y., YUAN L., HUA G.: Designing a better asymmetric vqgan for stablediffusion, 2023. 5
- [ZHY*23] ZHU J., HUO Y., YE Q., LUAN F., LI J., XI D., WANG L., TANG R., HUA W., BAO H., ET AL.: I2-sdf: Intrinsic indoor scene reconstruction and editing via raytracing in neural sdf. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog.* (2023). 3
- [ZIE*18] ZHANG R., ISOLA P., EFROS A. A., SHECHTMAN E., WANG O.: The unreasonable effectiveness of deep features as a perceptual metric. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog.* (2018). 10
- [ZLW*21] ZHANG K., LUAN F., WANG Q., BALA K., SNAVELY N.: PhySG: Inverse rendering with spherical gaussians for physics-based material editing and relighting. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog.* (2021). 3
- [ZLZ*22] ZHU Z.-L., LI Z., ZHANG R.-X., GUO C.-L., CHENG M.-M.: Designing an illumination-aware network for deep image relighting. *IEEE Trans. Image Process.* 31 (2022), 5396–5411. 2
- [ZRA23] ZHANG L., RAO A., AGRAWALA M.: Adding conditional control to text-to-image diffusion models. In *IEEE/CVF Int. Conf. Comput. Vis.* (2023). 2, 3, 4, 7
- [ZSD*21] ZHANG X., SRINIVASAN P. P., DENG B., DEBEVEC P., FREEMAN W. T., BARRON J. T.: Nerfactor: Neural factorization of shape and reflectance under an unknown illumination. *ACM Trans. Graph.* 40, 6 (2021), 1–18. 2, 3

[ZSH*22] ZHANG Y., SUN J., HE X., FU H., JIA R., ZHOU X.: Modeling indirect illumination for inverse rendering. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog.* (2022). 3

[ZTS*22] ZHANG X., TSENG N., SYED A., BHASIN R., JAIPURIA N.: Simbar: Single image-based scene relighting for effective data augmentation for automated driving vision tasks. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog.* (06 2022). 2

[ZYL*23] ZHANG J., YAO Y., LI S., LIU J., FANG T., MCKINNON D., TSIN Y., QUAN L.: Neilff++: Inter-reflectable light fields for geometry and material estimation. In *IEEE/CVF Int. Conf. Comput. Vis.* (2023). 3