Generative Multiview Relighting for 3D Reconstruction under Extreme Illumination Variation

Anonymous ICCV submission

Paper ID 0000

Figure 1. **3D reconstruction under extreme illumination variation.** We propose a method for 3D reconstruction from a set of images captured under strongly varying illumination. Our method recovers high-fidelity appearance details including specular highlights that prior state-of-the-art approaches cannot recover (*top baseline*: NeRF-Casting [47] with appearance embeddings, *bottom baseline*: NeROIC [25]).

Abstract

Reconstructing the geometry and appearance of objects from photographs taken in different environments is difficult as the illumination and, therefore, the object appearance vary across captured images. This is particularly challenging for specular objects whose appearance strongly depends on the viewing direction. Some prior approaches model appearance variation across images using a per-image embedding vector, while others use physically-based rendering to recover the materials and per-image illumination. Such approaches fail at faithfully recovering view-dependent appearance given the significant variation in input illumination and tend to produce mostly diffuse results. We present an approach that reconstructs objects from images taken under different illuminations by first relighting the images under a single reference illumination with a multiview relighting diffusion model and then reconstructing the object's geometry and appearance with a radiance field architecture that is robust to the minor remaining inconsistencies among the relit images. We validate our approach on synthetic and real datasets and demonstrate that it outperforms existing techniques at reconstructing high-fidelity appearance from images taken under extreme illumination variation. Moreover, our approach is particularly effective at recovering view-dependent "shiny" appearance which cannot be reconstructed by prior methods.

1. Introduction

A common strategy for performing novel view synthesis is to recover a 3D representation from a collection of photographs of a scene and then use it to render novel views from unseen viewpoints. Most view synthesis approaches assume that the input images are all taken under the same illumination conditions, *i.e.*, the scene's appearance is assumed to be *static*. However, this assumption is frequently violated during capture: moving clouds may cause the intensity of the sun to vary, artificial light sources may turn on or off, and even the shadows cast by the photographer may interfere with the scene. This issue becomes even more se-

vere when dealing with unstructured photograph collections such as those scraped from the Internet, where an object may be imaged in environments with extremely different illumination conditions, such as the case shown in Figure 1.

One strategy for addressing illumination variation is to use a view synthesis model capable of representing appearance changes across captured images. A simple and popular approach is to parameterize the view-dependent radiance of the scene as a function of a per-image "latent code" [9, 34]. Such an approach is effective for modeling changes (*i.e.* the "base" color and texture of the object). However, the additional expressivity comes at a cost because it tends to "explain away" *all* view-dependent appearances (not just those due to variable illumination). For non-diffuse (shiny) objects, we find that methods using per-image latent code often produce unrealistic, mostly-diffuse reconstructions.

Other physics-based approaches explicitly recover the spatially-varying material properties of the scene in addition to a per-image representation of illumination [10, 25]. While inverse rendering provides a more physically meaningful image formation model than a view-dependent latent encoding, it suffers from the same ambiguities: appearance variation can be attributed to either a change in lighting or a change in viewing direction, and these models often mix up the two.

Recent works [24, 58] attempt to circumvent these ambiguities by training image diffusion models that relight an individual image to appear as if it were lit by a specified illumination condition. Though diffusion models can leverage strong priors on natural image appearance to assist in material/lighting disambiguation, single-image relighting is fundamentally ill-posed. As such, independently applying these models to relight multiple images to a single reference illumination often yields samples corresponding to completely different, mutually-incompatible explanations of object materials.

In this work, we first use a *multiview* image diffusion model to jointly relight all input images to match the lighting of a reference image selected from the set of input images. This produces relit images that are significantly more consistent than those from prior single-image relighting diffusion models. We then design a novel 3D reconstruction model based on NeRF [35, 47] to fit a 3D representation to those relit images, which enables rendering from novel viewpoints. Our 3D reconstruction model is designed to be robust to inconsistencies that may still be present in the output of the relit samples. Since the diffusion model has to implicitly estimate 3D shape to relight the input images, any errors in its shape estimation would translate to specular highlights not appearing at the correct locations. To resolve this, we use a per-image vector which we call "shading embedding". This vector is used to encode per-image perturbations to the surface normal vectors used by the shading model of the NeRF, which allows the model to account for errors in the model's underlying estimated shape.

We validate the method on both synthetic (Objaverse [13]) and real (NAVI [22]) datasets. We showcase the capability of reconstructing accurate geometry and accurate view-dependent appearance from images captured under extreme illumination variation. Extensive quantitative and qualitative evaluation shows that the proposed approach significantly outperforms state-of-the-art methods both qualitatively and quantitatively.

2. Related Work

3D Reconstruction and View Synthesis. Neural radiance fields (NeRF) [35] is a successful approach for reconstructing a 3D scene under a fixed illumination. Many improvements to NeRF have been developed to improve the modeling of effects such as refractions and reflections [1, 6, 28, 32, 46, 47, 50], but all of these approaches also rely on static lighting. NeRF-W [34] achieved high-quality 3D reconstructions from unconstrained "in the wild" images taken from the internet, which tend to exhibit a variety of inconsistencies in lighting, appearance, and geometry. One aspect of this approach was using per-image appearance embeddings within the learned mapping from spatial coordinates and viewing direction to color (but not volumetric density). This parameterization forces optimization to recover a single consistent 3D geometry model while allowing appearance to vary across input images. This allows per-image appearance variation to be explained away, but also allows many view-dependent effects to be incorrectly attributed to per-image variation, often resulting in diffuse and erroneous reconstructions. Due to its simplicity (in that it does not contain a physically-based image formation model), per-image appearance embeddings are frequently used in settings where lighting only varies slightly [3, 4, 45].

Inverse Rendering. Decomposing an image into its constituent physical components has been a central problem in computer vision for nearly half a century [5]. Most modern approaches leverage the success of the computer graphics community, which has produced accurate and efficient models for the "forward" problem of rendering an image from an underlying 3D model [36]. Many modern inverse rendering techniques use radiance fields that, rather than mapping a 3D location and viewing direction to an outgoing color, map a 3D location to material properties and surface normals, which are then rendered according to some estimate of incident illumination [2, 7, 10, 11, 14, 23, 29–31, 33, 43, 44]. Because this problem is ill-posed, these techniques often critically rely on analytical priors to regularize the estimated scene decomposition. Although these models fully decompose the scene into its components and thereby enable editing and relighting, their rigidly specified

rendering models tend to constrain the solution space of the recovered model and result in lower rendering quality than models designed solely for view synthesis.

Many prior approaches to inverse rendering depend critically on modeling illumination and material properties using low-frequency parameterizations such as spherical harmonics [25], spherical Gaussians [55], or MLPs that are strongly biased to produce smoothly-varying outputs [57]. All of these parametrizations necessarily limit the model's ability to recover and render high-frequency (*i.e.*, non-smooth) color variations caused by glossy surfaces. This limitation is avoided by NeRF-Casting [47], which models incident illumination by explicitly reflecting secondary rays into the volumetric representation, thereby enabling the reconstruction of extremely specular objects. The 3D representation used by our model builds upon NeRF-Casting.

Lighting Estimation. Because estimating every aspect of a 3D scene is a challenging task with a broad scope, many prior works have instead attempted to only reconstruct a model of illumination from input images [16, 17, 26, 42]. Though this problem is still fundamentally under-constrained, it can be made tractable by learning a model from datasets containing light probes [27] or by relying on accidental light probes in the scene [51]. Modern generative approaches in this vein have shown that diffusion models are capable of developing an internal understanding of illumination [37], which is further validated by our work.

Generative Relighting. Diffusion-based generative models have recently been used for relighting individual images [52] and, closer to our approach, for 3D relighting [24, 38, 58]. This process circumvents the need to decompose the scene into physically-meaningful components and instead relies on a diffusion model to re-synthesize an input image as if it were illuminated by some reference illumination. Unlike our method, these approaches require the set of input images to have constant illumination, which is used in the first step to recover the 3D geometry of the object before relighting it. IllumiNeRF [58] samples a diffusion model multiple times to generate a collection of plausible results, while Neural Gaffer [24] iteratively refines a set of relit images alongside a NeRF. Crucially, both models *independently* relight each input image given a reference illumination, which can lead to inconsistent relit images. Concurrent to our work, RelitLRM [56] uses a large reconstruction model to jointly reconstruct and relight an object from multiple images, but it also assumes that the input images are consistently lit. Additionally, all of these methods require a reference environment map as input, while we address a more general problem where we only have access to images. Our model works by relighting a set of differently-lit input images *simultaneously* to match the appearance of one of the input images, and it does not require any additional information, including constantly-lit input images or a reference environment map. We show that our approach not only yields significantly more consistent 3D reconstructions than prior work, but is also more practically applicable to a broader variety of inputs and capture settings.

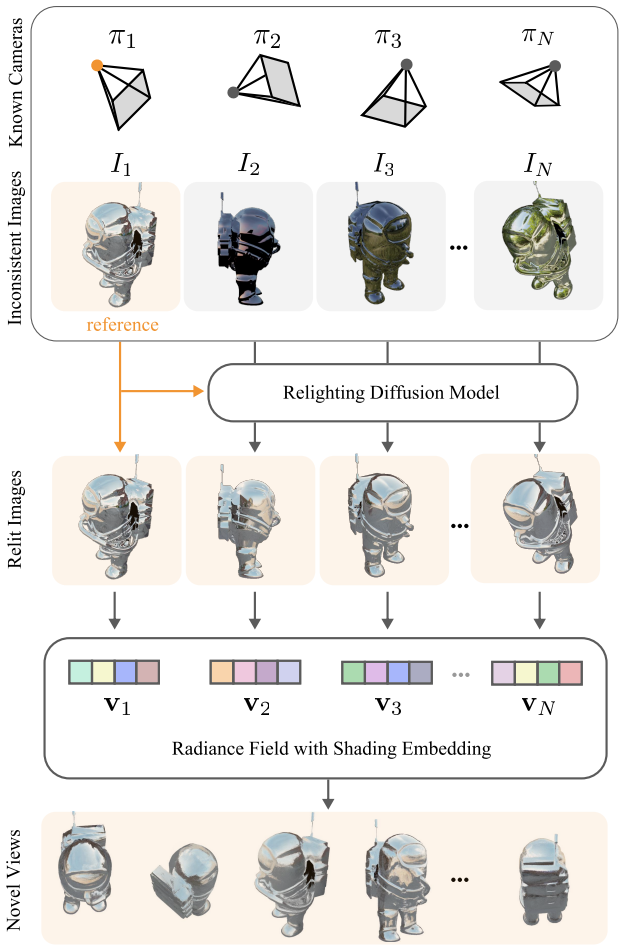


Figure 2. **Method overview.** We first apply a relighting diffusion model that converts N images I_1, ..., I_N with known camera poses π_1, ..., π_N, captured under extremely different illuminations, to a set of images with the same poses, but rendered under the illumination of the reference image I_1 (highlighted in orange). We then optimize a neural radiance field to obtain a consistent 3D representation with a novel per-image shading embedding, which can be used to render new views of the scene from unobserved poses.

3. Method

We aim to recover a 3D model from an collection of (posed) images of an object, where each image is illuminated by arbitrarily-varying lights. We use a two-step solution: in Section 3.1 we describe a relighting diffusion model that "harmonizes" the input images by making their illumina-

tion constant, and in Section 3.2, we describe a NeRF-based approach to convert the relit images produced by that diffusion model into a consistent 3D model. Our full pipeline is illustrated in Figure 2.

3.1. Relighting model

Our relighting model takes a set of N input images $I_1, ..., I_N$ of an object from multiple known camera poses $\pi_1, ..., \pi_N$, each of which is assumed to have been taken under a different illumination (see Figure 2, leftmost pane). The relighting model's aim is to generate a set of images of the same object from the same viewpoints, relit under the illumination from a designated reference image in the input set, which we arbitrarily select as the first image, I_1.

We build our relighting model using image diffusion, which can leverage strong natural image priors from large amounts of data. While recent works have used image diffusion models to relight individual input images based on a reference environment map [24, 52, 58], we argue that it is essential to relight the entire set of input images jointly. As demonstrated by the success of photometric stereo [19], leveraging multiple observations of the same material under varying lights reduces the inherent ambiguity between material properties and illumination, and it provides a strong cue for estimating shape. Using this same insight in the context of modern diffusion models leads to significantly more consistent relit images, as each sample from our model represents a unified interpretation of the object's material. Consequently, our method does not require ad-hoc solutions to address the issue of extremely inconsistent relit images, such as the use of appearance embeddings [58] or refinement of the relit images during reconstruction [24].

Our model's architecture follows prior work on view synthesis that extends image generative models to the multiview or video case [8, 15, 21] — we will analogously extend a single-image relighting model to multiple views, each with a distinct light.

We construct our relighting diffusion model as a multiview latent model that denoises a set of N latent codes $\mathbf{z}_1, ..., \mathbf{z}_N$, one for each input image. Unlike generative models that synthesize images from only random noise, the observed images $I_1, ..., I_N$ contain a significant amount of valuable information about geometry and illumination. As such, when denoising the ith latent code, $\mathbf{z}_i$, the denoising network should use the geometry encoded in image I_i as well as the lighting of the reference image I_1. To accomplish this, we pass the "clean" latent corresponding to the ith image I_i and an encoding of its pose π_i into the network that denoises the ith latent $\mathbf{z}_i$. To jointly reason about the reference lighting, we use 3D self-attention blocks [15] for the latent codes, and we use cross-attention in between them and the camera poses. This structure allows the geometry and camera pose of the ith image and the lighting of the reference image to be directly observed by the network denoising the ith latent $\mathbf{z}_i$.

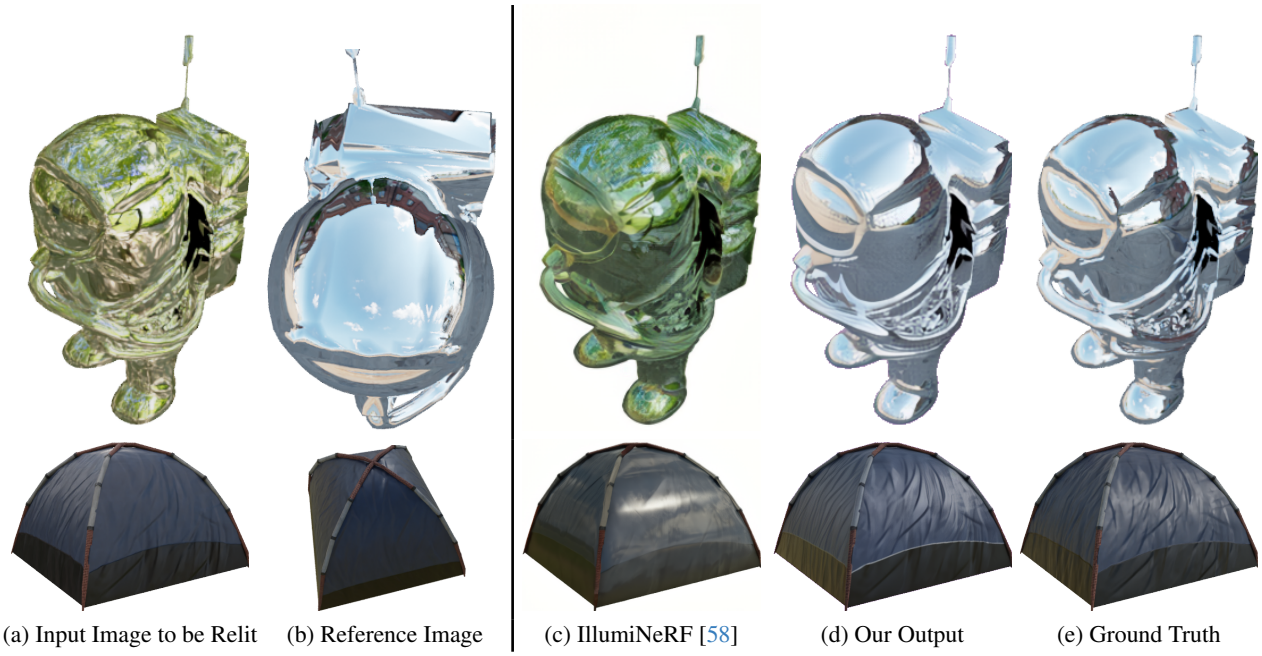
Because our input is an unordered set of images, we follow CAT3D's approach [15] of replacing temporal embeddings with camera poses which we encode as raymaps [41, 49]. To inform the network which image should be used as the reference, we concatenate the reference image with a "reference map" consisting of a single-channel image of ones, and we concatenate all other images with a single-channel image of zeros. See the supplement for a full description of this process. Results from this relighting model component are shown in Figure 3.

3.2. 3D Reconstruction

Applying the relighting model from Section 3.1 to the input images yields a set of "harmonized" input images that appear to have been lit under a single consistent illumination. With these images we can recover a 3D model without solving the difficult problem of handling variable illumination within a 3D reconstruction pipeline. We build our reconstruction model on NeRF-Casting [47], a view synthesis technique that achieves state-of-the-art results for reconstructing highly specular objects under constant illumination.

Although the relit images produced by our relighting diffusion model appear largely consistent when visually inspecting individual images, they still exhibit slight inconsistencies. As we will demonstrate, if the outputs of our diffusion model are used to naïvely train a NeRF that is not robust to this subtle variation (such as the common appearance-embedding based approach of NeRF-W [34]), these inconsistencies may be absorbed into the view-dependent parameterization of the NeRF, which results in novel view synthesis results with highly unrealistic and distracting "flickering" temporal artifacts as the camera moves. We must, therefore, be careful to parameterize appearance such that optimization is robust to this variation — resolving these inconsistencies is a primary goal of 3D reconstruction.

We address this problem by observing that the subtle inter-image errors made by the diffusion model are usually due to the specular highlights being slightly "warped" relative to the ground truth. This suggests that the diffusion model has small errors in its implicit estimate of the surface normals of the object, which causes the specular reflections to be tilted in slightly incorrect directions. We therefore resolve this by using a per-image "shading embedding": an embedding vector $\mathbf{v}_i$ is used to warp the normal vectors of the corresponding ith training image, and those warped normal vectors are provided as input to the network that models the shading (and therefore, appearance) of the image. More concretely, the normals of the ith image at a 3D coordinate



(a) Input Image to be Relit (b) Reference Image (c) IllumiNeRF [58] (d) Our Output (e) Ground Truth

Figure 3. **A comparison of our multiview relighting with prior work on single-image relighting.** Our method first relights a set of inconsistently-lit images (one of which is shown in (a)) to match the illumination of a selected reference image (b) in that set. Single-image relighting techniques such as IllumiNeRF [58] (c) struggle to disambiguate geometry, lighting, and materials, leading to an inaccurate relighting. In contrast, our model jointly relights a set of inconsistently-lit frames, which reduces ambiguities and results in a significantly more accurate result (d) when compared with the ground truth (e).

$\mathbf{x}$ are computed as:

$$\mathbf{n}_i(\mathbf{x}) = \text{normalize}(\text{MLP}(\mathbf{f}(\mathbf{x}), \mathbf{v}_i)) \,, \tag{1}$$

where $\text{MLP}(\cdot)$ is a multi-layer perceptron with 3 layers each with 128 hidden units, $\mathbf{f}(\mathbf{x})$ is a spatially-varying feature parameterizing the normals (identical to the one used by NeRF-Casting), and the $\text{normalize}(\cdot)$ operator turns the output of the MLP to a unit-length vector.

The per-image surface normals defined in Equation 1 can slightly alter the directions in which secondary rays are reflected by the model, which allows optimization to account for errors in the output of the relighting diffusion model. We prevent the normals from changing significantly between different images by tying them to the normals corresponding to the underlying NeRF density field, as in NeRF-Casting [47]. See NeRF-Casting and our supplement for more details. As we will demonstrate later, this shading embedding yields more accurate reconstructions and renderings than the standard appearance embedding approach [34] used in prior work for relighting [24, 58].

4. Architecture and Training

Training Data. The relighting model's training dataset is key to its ability to generalize across different illumination conditions, materials, and geometries. Due to the ease of rendering a large set of synthetic data, we rely solely on synthetic data for training, and we experimentally verify that our method generalizes to real captured photographs. We render our dataset using assets from a dataset of $\sim$ 300K high-quality objects, similar in appearance to Objaverse but with more diverse materials. We additionally augment this dataset by rendering another copy after replacing all materials with perfect mirrors, which we found improved reconstruction quality for both highly-specular objects and (perhaps surprisingly) for standard mostly-diffuse objects. We render the objects using about 700 environment maps taken from Poly Haven [39], which we augment by randomly rotating azimuthally.

Diffusion Model. We initialize our model using an image generation latent diffusion model similar in architecture to Stable Diffusion 1.5 [40] that had been trained on a large dataset of images. The image input of our diffusion model is a $512 \times 512 \times 3$ resolution image that is encoded into $64 \times 64 \times 8$ latents. To enable classifier-free guidance (CFG) [20], we randomly mask out the attention to the reference image to force the model to do "unconditional relighting" and relight the input views with an arbitrary lighting. For all our experiments, we use a CFG value of 3. Because the speed and memory requirements of training the model are inversely proportional to the number of synthesized views, we train our model in stages: We first train the

model to relight 8 frames for 200K steps, then 16 frames for 100K steps, then 32 frames for 50K steps, and finally 64 frames for another 50K steps. Note that our model parameters are independent of the number of input views, since the cross-attention operation can scale to an arbitrary number of images. Our model was trained on 64 5th-generation TPUs with a total batch size of 64, which took approximately two weeks. At inference, we set the number of images that our diffusion model simultaneously relights to $N = 64$

3D Reconstruction. The optimization of each radiance field uses a similar number of parameters to those used in NeRF-Casting [47]. Optimizing a model from a set of harmonized images takes roughly 30 minutes on 16 NVIDIA A100 GPUs, and rendering a single 512×512 image takes about 0.5 seconds on the same hardware. See supplement for details and a full description of all hyperparameters.

5. Results

We evaluate our method on two datasets: synthetic objects from Objaverse [13] and real captured objects from NAVI [22]. The synthetic dataset comprises two components: 8 standard textured assets whose materials range from mostly-diffuse to glossy, and 12 shiny assets containing objects with perfectly-reflective materials, designed to highlight the challenge of recovering accurate view-dependent effects. We render the synthetic Objaverse dataset using environment maps from HDRIMaps [18]. Each scene contains 64 training images and 36 testing images, all with different random poses. Each training image was rendered with distinct illumination, and the test set was rendered using the illumination of the reference image from the training set. Please see the supplemental material for additional results and videos.

5.1. Comparisons to State-of-the-Art Methods

We compare our approach with two methods designed for 3D reconstruction "in the wild": NeROIC [25] and NeRF-Casting [47] with standard appearance embeddings for each image (which we call "NeRF-Casting + AE" or "NeRFCast + AE" for short). We also compare our approach with the recent diffusion-based relighting method IllumiNeRF [58]. IllumiNeRF requires a target environment map for relighting and we provide it with the ground truth illumination, which is a significant advantage over other techniques as the full environment map contains more information than the reference image. For methods with appearance embeddings (ours and NeRF-Casting + AE), we render novel views using the same embedding as the reference image. Similarly, when rendering test images using NeROIC, we use the estimated illumination of the reference image.

Synthetic Results. In Figure 4, we qualitatively compare our method's novel view synthesis results with prior state-of-the-art approaches. Our results appear significantly more accurate, and all baselines struggle to recover specular highlights. This is particularly apparent in the first and third rows, demonstrating the recovery of fine details in the reflections of perfectly specular objects. The quantitative results in Table 1 show that our method significantly outperforms all baselines on both splits of the dataset, even when we provide the IllumiNeRF [58] baseline with true environment maps corresponding to the reference image.

Table 1. **Novel view relighting on Objaverse.** Our approach of relighting and robustly optimizing a radiance field outperforms previous state-of-the-art methods in rendering novel views under target lighting. We outperform prior work despite giving the baseline a significant advantage by providing it the true lighting.

	Standard Assets			Shiny Assets		
Method	**PSNR↑**	**SSIM↑**	**LPIPS↓**	**PSNR↑**	**SSIM↑**	**LPIPS↓**
NeROIC [25]	26.13	0.935	0.088	22.14	0.880	0.113
NeRFCast [47] + AE	27.53	0.941	0.067	21.80	0.874	0.108
IllumiNeRF [58] w/ known light	29.22	0.958	0.057	23.46	0.881	0.095
Ours	**31.34**	**0.966**	**0.053**	**26.54**	**0.911**	**0.090**

Real Photographs. The top row of Figure 1 shows photographs we captured of a chrome-plated figurine in different environments. The resulting synthesized novel views show that our method can render accurate images corresponding to the selected reference illumination. In the bottom of Figure 1 and in Figure 5 we show reconstruction results on captured photographs from NAVI. The results show that our method can preserve shadows and specular highlights. Since NAVI also provides multiview data for each illumination, we can quantitatively evaluate the novel view synthesis performance on held-out views under the same illumination as the reference. We evaluate the method on all of the 18 NAVI scenes with at least 64 frames, and we show that our method also achieves the best quantitative results across all prior methods, as shown in Table 2. Since NAVI does not provide ground truth environment maps, we could not evaluate IllumiNeRF on this dataset.

Table 2. **Novel view relighting on real photos from NAVI.** We show that we outperform prior work on synthesizing novel views from "in-the-wild" images.

Method	**PSNR↑**	**SSIM↑**	**LPIPS↓**
NeRF-Casting [47] + AE	23.59	0.916	0.066
NeROIC [25]	24.01	0.918	0.079
Our Model	**25.55**	**0.929**	**0.060**

5.2. Ablation Studies

Here, we provide ablations designed to test the different components of our model.

(a) Reference (b) Sample Inputs (c) NeRFCast + AE (d) NeROIC (e) IllumiNeRF (f) Ours (g) Ground Truth

Figure 4. **Visual comparison of novel view renderings on the Objaverse dataset**. (b) We show sample input images under extreme illumination variation. (c) Adding a per-image latent code to NeRF-Casting [47] ("NeRFCast + AE") cannot accurately explain away the variations, leading to erroneous reconstruction. (d) Due to the ill-posed nature of the problem, inverse rendering-based methods such as NeROIC [25] tend to produce lower-quality renderings with mostly diffuse appearance. (e) IllumiNeRF [58] leverages diffusion prior for single-image relighting but produces inconsistent output samples, resulting in excessive blur in rendered novel views. Note that IllumiNeRF requires access to the target illumination's environment map as input. We provide IllumiNeRF with the ground truth environment map corresponding to the reference image (a), while other methods only have access to the reference image itself. (f) Our method renders accurate appearance with specular highlights close to those in the ground truth images (g).

Number of Frames. A major benefit of our approach is that it jointly relights all frames. Table 3 shows our view synthesis results when relighting the images using different numbers of simultaneously-predicted frames N. Increasing the number of predicted frames improves 3D reconstruction, hence our choice for predicting $N = 64$ images.

Table 3. **The effect of the number of frames N output simultaneously by our model.** Increasing the number of frames simultaneously processed by our model improves its performance, therefore we use $N = 64$.

	Standard Assets			Shiny Assets		
# frames N	**PSNR↑**	**SSIM↑**	**LPIPS↓**	**PSNR↑**	**SSIM↑**	**LPIPS↓**
1 frame	28.39	0.943	0.079	23.85	0.889	0.105
8 frames	30.43	0.956	0.067	25.62	0.901	0.095
16 frames	30.49	0.958	0.063	25.52	0.900	0.097
32 frames	30.72	0.961	0.060	26.14	0.908	0.090
64 frames (ours)	**31.34**	**0.966**	**0.053**	**26.54**	**0.911**	**0.090**

Dataset. Next, we demonstrate the importance of including purely-specular objects in our training data. To do this, we train two 16-frame versions of our model for 70K steps, with and without the purely specular data augmentation described in Section 4. Table 4 shows that including the specular materials in the training set improves performance on specular objects and, perhaps surprisingly, that it is also beneficial for our general less-specular (but not perfectly-diffuse) dataset.

Table 4. **Training dataset ablation.** We show that training on assets with standard materials as well as objects with highly reflective materials improves performance for both types of assets.

	Standard Assets			Shiny Assets		
Training data	**PSNR↑**	**SSIM↑**	**LPIPS↓**	**PSNR↑**	**SSIM↑**	**LPIPS↓**
Standard assets	29.16	0.938	0.066	22.34	0.874	0.113
Standard + shiny assets	**29.26**	**0.939**	**0.066**	**22.91**	**0.877**	**0.111**

Shading Embeddings. Table 5 demonstrates the effect of our shading embeddings on the 3D reconstruction of the shiny synthetic scenes harmonized by our relighting model. As explained in Section 3.2, the shading embeddings allow the model to slightly shift the normal vectors used for computing appearance separately for each image, which enables absorbing small inconsistencies in the relit images. This improves upon the same NeRF model with no per-image em-

(a) Reference (b) Input Samples (c) NeRFCast + AE (d) NeROIC (e) Ours (f) Ground Truth

Figure 5. **Comparison on real world photos.** We use our method to reconstruct objects from in-the-wild photos taken in different environments. Our method can render novel views under the illumination conditions of any input image we select as the reference. Unlike prior work, our technique accurately preserves shadows (*e.g.* the bunny's ear shadow in the first row) and reflections (*e.g.* the box's handle in the second row and the car's specularities in the last row) that appear in the reference image we would like to match.

beddings ("No embeddings"). In contrast, using standard embeddings as in NeRF-W [34] ("Shading embeddings") has an adverse effect: the per-image codes allow the model to encode specularities as per-image diffuse color, which results in unrealistic and poor reconstructions.

Table 5. **Per-image embedding ablation.** We show that our shading embeddings improve performance compared to NeRF-Casting without any per-image code and that standard appearance embeddings have an adverse effect on reconstruction quality.

Method	PSNR↑	SSIM ↑	LPIPS↓
No embeddings	26.02	0.907	0.094
Appearance embeddings	24.38	0.896	0.096
Shading embeddings (ours)	**26.54**	**0.911**	**0.090**

6. Discussions and Conclusions

Limitations. Our approach requires object masks and accurate camera poses as inputs. Camera poses, in particular, can be challenging to compute from images of highly reflective objects that do not have reliable features for matching, but are required by most 3D reconstruction methods. However, there has been significant recent progress in camera pose estimation using learning-based techniques [12, 48, 53, 54] which do not rely on explicit feature matching. We believe that leveraging strong generative priors for joint pose estimation and 3D reconstruction is an exciting direction for future research.

Conclusions. In this paper, we address the challenge of reconstructing 3D objects from images captured under extreme illumination variation. Our core insight is to simultaneously relight all input images to match the illumination of a chosen reference image, thereby harmonizing lighting conditions among input images. We achieve this by training a diffusion-based multiview relighting model. Using these relit images, we apply a 3D reconstruction model that is robust to residual inconsistencies, enabling accurate reconstruction of the object's shape and view-dependent appearance. Our method effectively reconstructs objects with complex shapes and materials from images taken under drastically different lighting conditions. Our work highlights a promising direction for leveraging strong generative priors to tackle the ill-posed problems of 3D reconstruction.

References

[1] Benjamin Attal, Jia-Bin Huang, Michael Zollhöfer, Johannes Kopf, and Changil Kim. Learning neural light fields with ray-space embedding. *CVPR*, 2022. 2

[2] Benjamin Attal, Dor Verbin, Ben Mildenhall, Peter Hedman, Jonathan T. Barron, Matthew O'Toole, and Pratul P.

Srinivasan. Flash Cache: Reducing Bias in Radiance Cache Based Inverse Rendering. *ECCV*, 2024. 2

[3] Jonathan T. Barron, Ben Mildenhall, Dor Verbin, Pratul P. Srinivasan, and Peter Hedman. Mip-NeRF 360: Unbounded Anti-Aliased Neural Radiance Fields. *CVPR*, 2022. 2

[4] Jonathan T. Barron, Ben Mildenhall, Dor Verbin, Pratul P. Srinivasan, and Peter Hedman. Zip-NeRF: Anti-aliased grid-based neural radiance fields. *ICCV*, 2023. 2

[5] Harry Barrow, J. Tenenbaum, A. Hanson, and E. Riseman. Recovering intrinsic scene characteristics. *Computer Vision Systems*, 1978. 2

[6] Mojtaba Bemana, Karol Myszkowski, Jeppe Revall Frisvad, Hans-Peter Seidel, and Tobias Ritschel. Eikonal fields for refractive novel-view synthesis. *SIGGRAPH*, 2022. 2

[7] Sai Bi, Zexiang Xu, Pratul P. Srinivasan, Ben Mildenhall, Kalyan Sunkavalli, Miloš Hašan, Yannick Hold-Geoffroy, David Kriegman, and Ravi Ramamoorthi. Neural reflectance fields for appearance acquisition. *arXiv:2008.03824*, 2020. 2

[8] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. *CVPR*, 2023. 4

[9] Piotr Bojanowski, Armand Joulin, David Lopez-Pas, and Arthur Szlam. Optimizing the latent space of generative networks. *ICML*, 2018. 2

[10] Mark Boss, Raphael Braun, Varun Jampani, Jonathan T. Barron, Ce Liu, and Hendrik P.A. Lensch. NeRD: Neural Reflectance Decomposition from Image Collections. *ICCV*, 2021. 2

[11] Mark Boss, Varun Jampani, Raphael Braun, Ce Liu, Jonathan T. Barron, and Hendrik P.A. Lensch. Neural-PIL: Neural Pre-Integrated Lighting for Reflectance Decomposition. *NeurIPS*, 2021. 2

[12] Eric Brachmann, Jamie Wynn, Shuai Chen, Tommaso Cavallari, Áron Monszpart, Daniyar Turmukhambetov, and Victor Adrian Prisacariu. Scene coordinate reconstruction: Posing of image collections via incremental learning of a relocalizer. In *ECCV*, 2024. 8

[13] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. *CVPR*, 2023. 2, 6

[14] Andreas Engelhardt, Amit Raj, Mark Boss, Yunzhi Zhang, Abhishek Kar, Yuanzhen Li, Deqing Sun, Jonathan T. Barron, Hendrik P.A. Lensch, and Varun Jampani. SHINOBI: Shape and Illumination using Neural Object decomposition via BRDF optimization In-the-wild. *CVPR*, 2024. 2

[15] Ruiqi Gao, Aleksander Holynski, Philipp Henzler, Arthur Brussee, Ricardo Martin-Brualla, Pratul P. Srinivasan, Jonathan T. Barron, and Ben Poole. CAT3D: Create anything in 3D with multi-view diffusion models. *NeurIPS*, 2024. 4

[16] Marc-Andre Gardner, Kalyan Sunkavalli, Ersin Yumer, Xiaohui Shen, Emiliano Gambaretto, Christian Gagne, and Jean-Francois Lalonde. Learning to predict indoor illumination from a single image. *SIGGRAPH Asia*, 2017. 3

[17] Mathieu Garon, Kalyan Sunkavalli, Sunil Hadap, Nathan Carr, and Jean-Francois Lalonde. Fast spatially-varying indoor lighting estimation. *CVPR*, 2019. 3

[18] HDR Maps. https://hdrmaps.com. 6

[19] Carlos Hernández, George Vogiatzis, and Roberto Cipolla. Multiview photometric stereo. *IEEE TPAMI*, 2008. 4

[20] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv:2207.12598*, 2022. 5

[21] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *NeurIPS*, 2022. 4

[22] Varun Jampani, Kevis-Kokitsi Maninis, Andreas Engelhardt, Arjun Karpur, Karen Truong, Kyle Sargent, Stefan Popov, Andre Araujo, Ricardo Martin-Brualla, Kaushal Patel, Daniel Vlasic, Vittorio Ferrari, Ameesh Makadia, Ce Liu, Yuanzhen Li, and Howard Zhou. NAVI: Category-agnostic image collections with high-quality 3D shape and pose annotations. *NeurIPS*, 2023. 2, 6

[23] Haian Jin, Isabella Liu, Peijia Xu, Xiaoshuai Zhang, Songfang Han, Sai Bi, Xiaowei Zhou, Zexiang Xu, and Hao Su. TensoIR: Tensorial inverse rendering. *CVPR*, 2023. 2

[24] Haian Jin, Yuan Li, Fujun Luan, Yuanbo Xiangli, Sai Bi, Kai Zhang, Zexiang Xu, Jin Sun, and Noah Snavely. Neural gaffer: Relighting any object via diffusion. *NeurIPS*, 2024. 2, 3, 4, 5

[25] Zhengfei Kuang, Kyle Olszewski, Menglei Chai, Zeng Huang, Panos Achlioptas, and Sergey Tulyakov. NeROIC: Neural Rendering of Objects from Online Image Collections. *SIGGRAPH*, 2022. 1, 2, 3, 6, 7

[26] Jean-François Lalonde, Alexei A Efros, and Srinivasa G Narasimhan. Estimating the natural illumination conditions from a single outdoor image. *IJCV*, 2012. 3

[27] Chloe LeGendre, Wan-Chun Ma, Graham Fyffe, John Flynn, Laurent Charbonnel, Jay Busch, and Paul Debevec. Deeplight: Learning illumination for unconstrained mobile mixed reality. *CVPR*, 2019. 3

[28] Ruofan Liang, Huiting Chen, Chunlin Li, Fan Chen, Selvakumar Panneer, and Nandita Vijaykumar. ENVIDR: Implicit differentiable renderer with neural environment lighting. *ICCV*, 2023. 2

[29] Zhi-Hao Lin, Jia-Bin Huang, Zhengqin Li, Zhao Dong, Christian Richardt, Tuotuo Li, Michael Zollhöfer, Johannes Kopf, Shenlong Wang, and Changil Kim. Iris: Inverse rendering of indoor scenes from low dynamic range images. *CVPR*, 2025. 2

[30] Zhi-Hao Lin, Bohan Liu, Yi-Ting Chen, Kuan-Sheng Chen, David Forsyth, Jia-Bin Huang, Anand Bhattad, and Shenlong Wang. Urbanir: Large-scale urban scene inverse rendering from a single video. *3DV*, 2025.

[31] Yuan Liu, Peng Wang, Cheng Lin, Xiaoxiao Long, Jiepeng Wang, Lingjie Liu, Taku Komura, and Wenping Wang. NeRO: Neural Geometry and BRDF Reconstruction of Reflective Objects from Multiview Images. *SIGGRAPH*, 2023. 2

[32] Li Ma, Vasu Agrawal, Haithem Turki, Changil Kim, Chen Gao, Pedro Sander, Michael Zollhöfer, and Christian Richardt. SpecNeRF: Gaussian directional encoding for specular reflections. *CVPR*, 2024. 2

[33] Alexander Mai, Dor Verbin, Falko Kuester, and Sara Fridovich-Keil. Neural microfacet fields for inverse rendering. *ICCV*, 2023. 2

[34] Ricardo Martin-Brualla, Noha Radwan, Mehdi S. M. Sajjadi, Jonathan T. Barron, Alexey Dosovitskiy, and Daniel Duckworth. NeRF in the Wild: Neural Radiance Fields for Unconstrained Photo Collections. *CVPR*, 2021. 2, 4, 5, 8

[35] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. *ECCV*, 2020. 2

[36] Matt Pharr, Wenzel Jakob, and Greg Humphreys. *Physically based rendering: From theory to implementation*. MIT Press, 2023. 2

[37] Pakkapon Phongthawee, Worameth Chinchuthakun, Nontaphat Sinsunthithet, Amit Raj, Varun Jampani, Pramook Khungurn, and Supasorn Suwajanakorn. DiffusionLight: Light probes for free by painting a chrome ball. *CVPR*, 2023. 3

[38] Yohan Poirier-Ginter, Alban Gauthier, Julien Philip, Jean-François Lalonde, and George Drettakis. A Diffusion Approach to Radiance Field Relighting using Multi-Illumination Synthesis. *Computer Graphics Forum*, 2024. 3

[39] Poly Haven. https://polyhaven.com. 5

[40] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. *CVPR*, 2022. 5

[41] Mehdi SM Sajjadi, Henning Meyer, Etienne Pot, Urs Bergmann, Klaus Greff, Noha Radwan, Suhani Vora, Mario Lučić, Daniel Duckworth, Alexey Dosovitskiy, et al. Scene representation transformer: Geometry-free novel view synthesis through set-latent scene representations. *CVPR*, 2022. 4

[42] Pratul P. Srinivasan, Ben Mildenhall, Matthew Tancik, Jonathan T. Barron, Richard Tucker, and Noah Snavely. Lighthouse: Predicting lighting volumes for spatially-coherent illumination. *CVPR*, 2020. 3

[43] Pratul P. Srinivasan, Boyang Deng, Xiuming Zhang, Matthew Tancik, Ben Mildenhall, and Jonathan T. Barron. NeRV: Neural reflectance and visibility fields for relighting and view synthesis. *CVPR*, 2021. 2

[44] Cheng Sun, Guangyan Cai, Zhengqin Li, Kai Yan, Cheng Zhang, Carl Marshall, Jia-Bin Huang, Shuang Zhao, and Zhao Dong. Neural-pbir reconstruction of shape, material, and illumination. *ICCV*, 2023. 2

[45] Matthew Tancik, Vincent Casser, Xinchen Yan, Sabeek Pradhan, Ben Mildenhall, Pratul P. Srinivasan, Jonathan T. Barron, and Henrik Kretzschmar. Block-NeRF: Scalable Large Scene Neural View Synthesis. *CVPR*, 2022. 2

[46] Dor Verbin, Peter Hedman, Ben Mildenhall, Todd Zickler, Jonathan T. Barron, and Pratul P. Srinivasan. Ref-NeRF: Structured view-dependent appearance for neural radiance fields. *CVPR*, 2022. 2

[47] Dor Verbin, Pratul P. Srinivasan, Peter Hedman, Ben Mildenhall, Benjamin Attal, Richard Szeliski, and Jonathan T. Barron. NeRF-Casting: Improved View-Dependent Appearance with Consistent Reflections. *SIGGRAPH Asia*, 2024. 1, 2, 3, 4, 5, 6, 7

[48] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. DUSt3R: Geometric 3D vision made easy. *CVPR*, 2024. 8

[49] Daniel Watson, William Chan, Ricardo Martin-Brualla, Jonathan Ho, Andrea Tagliasacchi, and Mohammad Norouzi. Novel view synthesis with diffusion models. *arXiv:2210.04628*, 2022. 4

[50] Liwen Wu, Sai Bi, Zexiang Xu, Fujun Luan, Kai Zhang, Iliyan Georgiev, Kalyan Sunkavalli, and Ravi Ramamoorthi. Neural directional encoding for efficient and accurate view-dependent appearance modeling. *CVPR*, 2024. 2

[51] Hong-Xing Yu, Samir Agarwala, Charles Herrmann, Richard Szeliski, Noah Snavely, Jiajun Wu, and Deqing Sun. Accidental Light Probes. *CVPR*, 2023. 3

[52] Chong Zeng, Yue Dong, Pieter Peers, Youkang Kong, Hongzhi Wu, and Xin Tong. DiLightNet: Fine-grained lighting control for diffusion-based image generation. *SIGGRAPH*, 2024. 3, 4

[53] Junyi Zhang, Charles Herrmann, Junhwa Hur, Varun Jampani, Trevor Darrell, Forrester Cole, Deqing Sun, and Ming-Hsuan Yang. Monst3r: A simple approach for estimating geometry in the presence of motion. *arXiv preprint arXiv:2410.03825*, 2024. 8

[54] Jason Y Zhang, Amy Lin, Moneish Kumar, Tzu-Hsuan Yang, Deva Ramanan, and Shubham Tulsiani. Cameras as rays: Pose estimation via ray diffusion. *ICLR*, 2024. 8

[55] Kai Zhang, Fujun Luan, Qianqian Wang, Kavita Bala, and Noah Snavely. PhySG: Inverse rendering with spherical gaussians for physics-based material editing and relighting. *CVPR*, 2021. 3

[56] Tianyuan Zhang, Zhengfei Kuang, Haian Jin, Zexiang Xu, Sai Bi, Hao Tan, He Zhang, Yiwei Hu, Milos Hasan, William T. Freeman, Kai Zhang, and Fujun Luan. RelitLRM: Generative relightable radiance for large reconstruction models. In *The Thirteenth International Conference on Learning Representations*, 2025. 3

[57] Xiuming Zhang, Pratul P. Srinivasan, Boyang Deng, Paul Debevec, William T. Freeman, and Jonathan T. Barron. NeRFactor: Neural Factorization of Shape and Reflectance Under an Unknown Illumination. *SIGGRAPH Asia*, 2021. 3

[58] Xiaoming Zhao, Pratul P. Srinivasan, Dor Verbin, Keunhong Park, Ricardo Martin Brualla, and Philipp Henzler. IllumiNeRF: 3D Relighting Without Inverse Rendering. *NeurIPS*, 2024. 2, 3, 4, 5, 6, 7