LightSwitch: Multi-view Relighting with Material-guided Diffusion

Yehonathan Litman Fernando De la Torre Shubham Tulsiani
Carnegie Mellon University

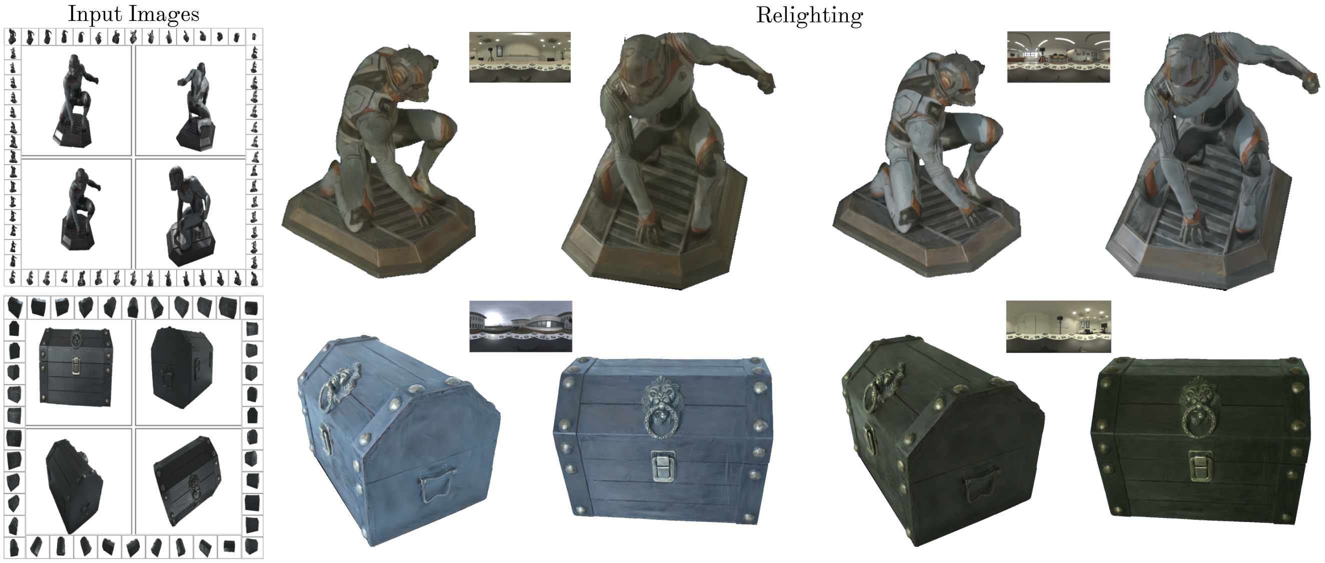

Figure 1. **Consistent Multi-view via Material-Guided Relighting Diffusion.** We present LightSwitch, a framework for multi-view relighting. Given any number of input images under an unknown illumination, LightSwitch leverages multi-view attention and inferred material properties to predict consistent relighting, enabling applications for 2D and 3D relighting.

Abstract

Recent approaches for 3D relighting have shown promise in integrating 2D image relighting generative priors to alter the appearance of a 3D representation while preserving the underlying structure. Nevertheless, generative priors used for 2D relighting that directly relight from an input image do not take advantage of intrinsic properties of the subject that can be inferred or cannot consider multi-view data at scale, leading to subpar relighting. In this paper, we propose Lightswitch, a novel finetuned material-relighting diffusion framework that efficiently relights an arbitrary number of input images to a target lighting condition while incorporating cues from inferred intrinsic properties. By using multi-view and material information cues together with a scalable denoising scheme, our method consistently and efficiently relights dense multi-view data of objects with diverse material compositions. We show that our 2D relighting prediction quality exceeds previous state-of-the-art relighting priors that directly relight from images. We further demonstrate that LightSwitch matches or outperforms state-of-the-art diffusion inverse rendering methods in relighting synthetic and real objects in as little as 2 minutes.

1. Introduction

We have witnessed impressive advances in the task of recovering 3D representations from multi-view captures, with methods like NeRF [27] and Gaussian Splatting [18] allowing one to reconstruct generic objects or scenes easily. While these representations excel at modeling detailed geometry and appearance, they only seek to model the appearance within the capture environment, thus baking in lighting effects into the obtained representation. This prevents reconstructions from being imported and relit in novel environments for applications such as virtual reality or synthesizing visual effects. In this work, we aim to enable such relightable rendering of 3D representations under generic illumination and present an approach that given posed multi-view images, enables synthesizing novel relit views.

Current methods that tackle this relighting task can be categorized as either leveraging inverse rendering or learned relighting. In particular, the former class of methods [10, 12, 16, 22, 23, 46, 48] seek to infer a 3D representation disentangling geometry, appearance, and material properties, allowing relighting via physics-based rendering. However, these optimization-based approaches tend to be slow and their usage of (simple) differentiable renderers can limit their ability to model complex lighting effects. In contrast, 'direct relighting' methods learn to directly generate a relit image a given (captured/rendered) source image and a target illumination. By adapting image diffusion priors, these methods can efficiently synthesize photo-realistic quality output in a feed-forward manner. However, these approaches operate on a single view, leading to inconsistencies in relighting across viewpoints.

Our work also adopts a learning-based approach for direct relighting, with the key insight that instead of formulating this as a single-view relighting task, we should formulate it as one of consistently relighting *multiple* input views. This can make the predictions consistent across views (making such a system better suited for 3D relighting), while also improving relighting performance as the cues observed in one view (*e.g.* sharpness of a specularity) can inform the relighting in another. In addition to incorporating multi-view cues, we also draw inspiration from inverse rendering methods that benefit from understanding material properties and seek to leverage (predicted) material properties (intrinsics and albedo) as additional input.

We build on these insights and propose 'LightSwitch', a relighting diffusion framework that produces multi-view consistent relighting informed by inferred intrinsic properties. We validate LightSwitch on both synthetic and real-world data and find that leveraging multi-view and predict material cues yield significantly improved relighting compared to prior learning-based relighting methods. We also tackle the '3D relighting' task requiring synthesizing and consistently relighting a large set of query views, and design a distributed inference scheme for efficient inference. We show that when compared to state-of-the-art inverse rendering methods, LightSwitch allows improved/comparable relighting, while being significantly faster.

2. Related Work

Image-based 3D Reconstruction. We have witnessed impressive recent progress in the task of recovering 3D from images. In particular representations like NeRF [27], Gaussian Splatting [18], and their variants [2, 3, 13, 26, 28, 35, 41] allow representing detailed geometry and appearance, and can be inferred from multi-view images of generic objects and scenes. More recent approaches, often leveraging generative priors, even allow inferring such detailed outputs from sparse or single-view input [7, 24, 25, 36, 39, 49], allowing end users to easily capture 3D. While these obtained reconstructs can capture the rich details of underlying 3D scenes, they are only capable of modeling the static environment observed in the images and cannot be imported to novel environments where their appearance would change under a different environmental illumination.

Relighting via Inverse Rendering. To enable relighting under novel environments, some works utilize inverse rendering to recover a 3D relightable asset. These approaches model intrinsic properties [10, 12, 16, 22, 23, 29, 45, 46] or light transport effects [5, 33, 48] to infer a factorized 3D representation that explains the observed images, allowing for relighting under novel illuminations. However, recovering intrinsics from a set of multi-view images is a non-trivial task, given that many different combinations of the intrinsic properties can be composed together to get the appearance in the source images. Some approaches use data-driven prediction of intrinsics [19, 40] to aid this, for example MaterialFusion [23], whose material model we adapt that uses a material prior model to aid with relighting. Nevertheless, inverse rendering pipelines are forced to rely on simple material models together with simpler lightweight differentiable renderers due to the computational constraints of physically-based renderers. Even so, optimization is still slow as real-time rendering requires considerable computation, making inverse rendering approaches time-consuming.

Learning Direct Relighting. An alternate relighting approach that allows for high fidelity relighting is to learn a model that directly predicts relit images, in particular by adapting generative diffusion priors for high quality generation. These direct feed-forward approaches allow for predicting accurate relighting in little time for high resolution images while generalizing to unseen instances [17, 42, 47]. However, not incorporating cues about the underlying material composition means the model is not using information that can help accurately relight assets with complex appearance effects. Even so, direct relighting models only take single-view images, limiting their practicality for 3D relighting where multi-view data provides important cues on the asset's inherent properties.

In contrast to aforementioned works, our proposed relighting diffusion framework incorporates intrinsic and multi-view cues to efficiently produce a high quality relighting directly from input images. The addition of these components to the relighting diffusion model significantly improves relighting for 2D and 3D scenarios. While some concurrent works have examined the task of recovering multi-view consistent relighting, they either do not leverage the prior knowledge of diffusion models [44], rely on multi-illuminated data [1], or incorporate material information cues for video relighting only [9, 21].

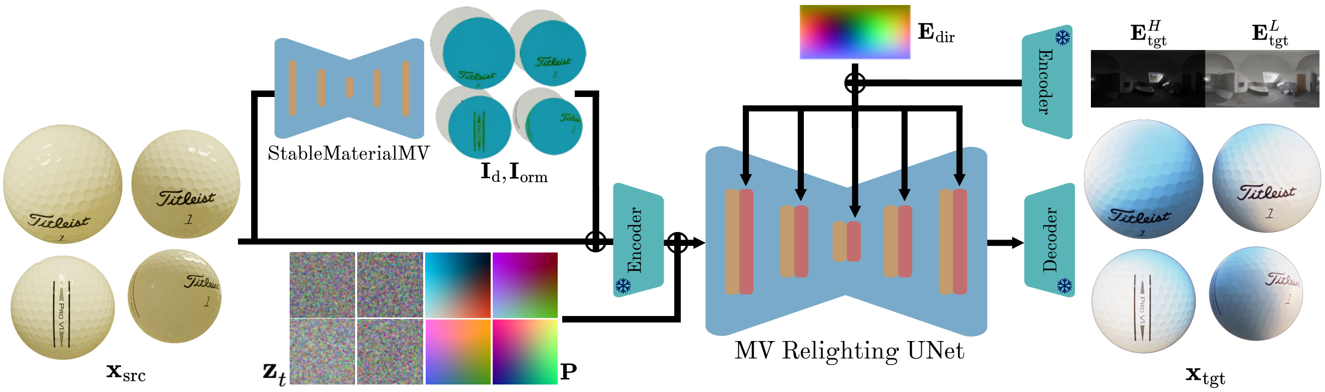
StableMaterialMV
$\mathbf{I}_\mathrm{d}, \mathbf{I}_\mathrm{orm}$
$\mathbf{E}_\mathrm{dir}$
$\mathbf{E}^H_\mathrm{tgt}$
$\mathbf{E}^L_\mathrm{tgt}$
Encoder
Encoder
Decoder
MV Relighting UNet
$\mathbf{x}_\mathrm{src}$
$\mathbf{z}_t$
$\mathbf{P}$
$\mathbf{x}_\mathrm{tgt}$

Figure 2. **LightSwitch Material-Relighting Diffusion Framework.** LightSwitch relights multi-view posed input images to a given target illumination. It infers and encodes multi-view consistent material image maps ($\mathbf{I}_\mathrm{d}, \mathbf{I}_\mathrm{orm}$) using a material diffusion model (StableMaterialMV [23]) and concatenates them to the Plücker ray maps ($\mathbf{P}$), encoded input images ($\mathbf{x}_\mathrm{src}$), and noisy latents ($\mathbf{z}_t$) in the channel dimension. The multi-view relighting UNet denoises the noisy latents and cross-attends to the lighting latents concatenated with the latent lighting directions ($\mathbf{E}_\mathrm{dir}$). The lighting latents are encoded from the processed target environment map images ($\mathbf{E}^H_\mathrm{tgt}, \mathbf{E}^L_\mathrm{tgt}$). The Stable Diffusion encoders and decoder are kept frozen.

3. Methodology

In this section we introduce our diffusion framework, shown in Fig. 2, that takes in multi-view posed input images captured under fixed unknown illumination and relights them to a given target illumination. We adapt a text-conditioned diffusion model and finetune it for multi-view material-guided relighting and describe this in Sec. 3.1. We then introduce an efficient relighting denoising scheme at inference for 3D relighting in Sec. 3.2, shown in Fig. 3.

3.1. Relighting Architecture

Our goal is to design an architecture that can consistently relight a set of input images given a target environment map. Towards this, we seek to adapt the deep image priors contained in large scale diffusion models like Stable Diffusion 2.1 [30]. Unlike previous approaches which focus on single-view relighting, we propose a model that allows multi-view consistent relighting. Moreover, in addition to conditioning on target illumination information, we also rely on material information cues for better relighting performance. We train this material-guided relighting diffusion model in stages beginning from single-view to yield a relighting model that synthesizes high quality multi-view consistent relightings.

Material Aware Single-view Relighting. The relighting diffusion UNet, initialized from Stable Diffusion 2.1's UNet, is first finetuned to relight single RGB input views captured under unknown lighting $\mathbf{x}_\mathrm{src}$ to target images $\mathbf{x}_\mathrm{tgt}$ illuminated by a given target lighting $\mathbf{E}_\mathrm{tgt}$. We modify the UNet's input layer to condition it on input images, material intrinsics, and camera pose information encoded as Plücker coordinate ray maps $\mathbf{P}$. To incorporate intrinsic material cues, the model utilizes per-pixel image material maps $\mathbf{I}_\mathrm{d}, \mathbf{I}_\mathrm{orm}$ corresponding to the albedo, occlusion, roughness, and metallicness (ORM)[1] components of the rendered image. The material representation used by the relighting model follows a simplified Disney principled BRDF model [6], where each pixel in $\mathbf{I}_\mathrm{d}, \mathbf{I}_\mathrm{orm}$ contain an albedo $\mathbf{a} \in \mathbb{R}^{H\times W\times 3}$, roughness $r \in \mathbb{R}^{H\times W\times 1}$, and metallicness $m \in \mathbb{R}^{H\times W\times 1}$. The underlying material information is constant across illuminations and using it as conditioning lets the relighting model understand how to relight views with diverse appearance effects such as specularities and absorption.

Incorporating Lighting. To incorporate lighting information into the denoising UNet, we add a lighting cross-attention module that attends to illumination information and finetune it together with the rest of the UNet. The target illumination information $\mathbf{E}_\mathrm{tgt}$ is given initially as a high dynamic range image and transformed to two environment maps following [17]: $\mathbf{E}^H_\mathrm{tgt}$, which is the normalized environment map, and $\mathbf{E}^L_\mathrm{tgt}$ which is tonemapped. The combination of these two maps helps inform the network about strong (e.g. lights, sun, etc.) and softer lighting (e.g. reflections, ambient lighting, etc.). These images are encoded using the Stable Diffusion encoder $\mathcal{E}$ and a directional embedding map $\mathbf{E}_\mathrm{dir}$ is also produced corresponding to the per latent lighting direction. The lighting module attends to the concatenation $(\mathcal{E}(\mathbf{E}^H_\mathrm{tgt}), \mathcal{E}(\mathbf{E}^L_\mathrm{tgt}), \mathbf{E}_\mathrm{dir})$.

At each training iteration we sample $(\mathbf{x}_\mathrm{src}, \mathbf{x}_\mathrm{tgt}, \mathbf{I}_\mathrm{d}, \mathbf{I}_\mathrm{orm}, \mathbf{P}, \mathbf{E}_\mathrm{tgt})$ and finetune the diffusion

[1] The occlusion map is not used and set to zero.

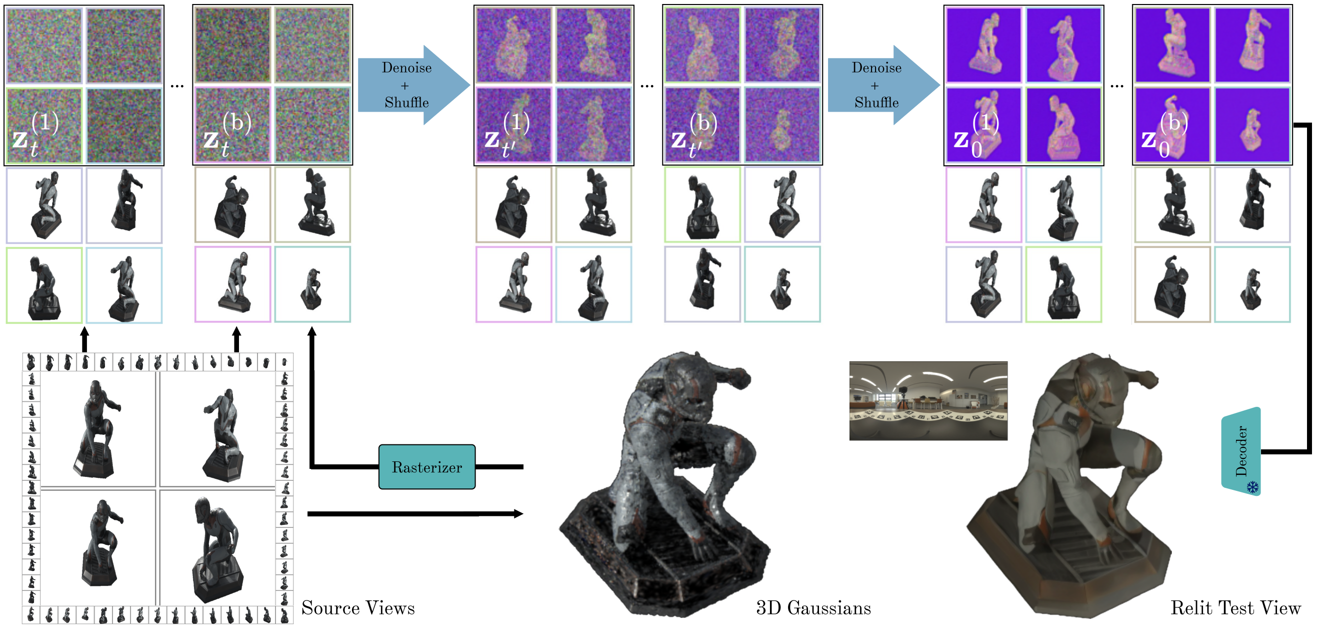

Figure 3. **Denoising Relighting for Any Number of Views.** Given the quadratic complexity of all-pair multi-view attention, we divide the input latents $\mathbf{z}_t$ into mini-batches $\mathbf{z}_t^{(1)}, \dots, \mathbf{z}_t^{(b)}$ and make latents attend to each other only within a subset per denoising iteration. When the batches are shuffled after the denoising step, they can attend to another subset in the next iteration. By continuously shuffling the subsets across DDPM iterations we approximate the full relighting diffusion. To relight a novel view, we optimize a 3D gaussian splat on the training images and render the novel view using the rasterizer. The novel view is then inserted into the set source views, enabling consistent novel view relighting.

model to denoise the noisy target view latent code $\mathbf{z}_t$. $\mathbf{z}_t$ is obtained by sampling a diffusion timestep t and a 4-channel random noise ϵ and adding the noise to $\mathcal{E}(\mathbf{x}_{\text{tgt}})$. We concatenate $(\mathcal{E}(\mathbf{x}_{\text{src}}), \mathcal{E}(\mathbf{I}_{\text{d}}), \mathcal{E}(\mathbf{I}_{\text{orm}}), \mathbf{P})$, denoted as $\mathbf{C}$, along the channel dimension of $\mathbf{z}_t$ as conditioning and process $\mathbf{E}_{\text{tgt}}$ for cross attention. The diffusion loss for the single-view training stage is

$$\mathcal{L}_{\text{diff}} = \|\epsilon_\theta\left(\mathbf{z}_t, t; \mathbf{C}\right) - \mathbf{v}_t\|_2^2, \tag{1}$$

where $\mathbf{v}_t$ corresponds to v-prediction loss [31].

Denoising Multi-view Relighting. After finetuning our relighting diffusion model to relight single-view posed images to a target illumination given groundtruth material information, we continue finetuning it for multi-view prediction. We modify the denoising UNet with multi-view self-attention modules and continue training. By first training the model for single-view relighting, it forms an initial understanding of lighting interaction between the environment and object that is considerably harder to model in multi-view. Previous work [32, 37] has shown that multi-view attention can be easily implemented in the self-attention module to enable consistent prediction across multiple views. Given a batch containing k image latents, the self-attention layer consolidates the batch together such that every latent pixel across the batch is in the same space and can attend to all other latent pixels. When integrating this into the relighting diffusion model, it predicts a consistent most probable relighting for a given illumination across all input appearance information. This is replicated for the material diffusion model for predicting the most probable material maps across given input views. We show that including the multi-view attention module significantly boosts relighting quality.

We sample k source views per object and use the following diffusion loss

$$\mathcal{L}_{\text{diff}}^{\text{mv}} = \|\epsilon_\theta\left(\mathbf{z}_t^{1:k}, t; \mathbf{C}^{1:k}\right) - \mathbf{v}_t\|_2^2, \tag{2}$$

We enable classifier-free guidance by setting $\mathcal{E}(\mathbf{E}_{\text{tgt}}^H), \mathcal{E}(\mathbf{E}_{\text{tgt}}^L)$ to all zeros with a 10% probability and set the guidance scale to 3. k is set to 4 during training. Once the base multi-view model is trained, we continue training an upscaled model at a resolution of 512×512 to produce higher fidelity relighting.

3.2. Lightswitch – Scalable Efficient 3D Relighting

With the trained upscaled multi-view LightSwitch diffusion model, we wish to apply it to 3D novel view relighting given sparse or dense multi-view data. In novel view relighting, training views of an asset are given as input and we wish to render an unseen view under a desired novel illumination. This can be practically challenging if the number of input views is too high as the compute requirement of transformers scales quadratically, and batch-wise processing can lead to inconsistencies. To address this, we introduce an efficient denoising mechanism that scales to an arbitrary number of input views as shown in Fig. 3. To enable novel relighting, we first render novel views under source illumination using novel view synthesis. We then input both the source views and synthesized novel views to our relighting network, allowing it to relight query views while using cues from the observed source views.

Distributed Multi-view Relighting. At inference time, we can effectively denoise more images by shuffling the data and sampling new batches per denoising iteration. Over enough iterations, each latent attends to all other latents across the entire dataset, making the final prediction consistent throughout. Additionally, we can distribute the denoising step in parallel across compute to proportionally accelerate the diffusion process. This keeps the consistency of the final prediction and allows for fast scalable relighting on high resolution images, enabling highly accurate relighting as seen in the example output in Fig. 1. The shuffling and denoising procedure is applied to both the material and relighting diffusion models for efficient material and relighting inference.

Rendering Test Views. Our approach distributes multi-view relighting effectively but can only relight given views. Thus, to relight novel views not included in the data, we optimize a 3D gaussian splat [18] on the training data, render and encode the test view, and include it in the training data being denoised. Given the strong performance of novel view synthesis approaches we can sample a high quality test view and easily relight it with LightSwitch, enabling low latency novel view relighting.

4. Experiments

We evaluate LightSwitch on image sets of diverse assets captured under varied illumination to showcase our approach's performance in 2D and 3D scenarios. To show LightSwitch's relighting performance and effective denoising scheme in 2D, we conduct a comparison against other diffusion-based relighting priors on a held out synthetic object test dataset relit with unseen target illuminations. We then evaluate our 3D novel view relighting method on both synthetic and real objects to highlight its generalization, efficiency, and relighting capabilities. This is compared against inverse rendering methods that also enable novel view relighting. Lastly, we ablate our LightSwitch to showcase how the choice of integrating material and multi-view information aids relighting.

4.1. Experimental Setup

Dataset. We curate a dataset of ~100K objects from a mixture of the BlenderVault [23] and Objaverse [8, 14] data that was filtered to include high quality objects containing PBR maps. We sample 8 camera poses on a hemisphere around an object and render those views under 8 different environment maps. The environment maps are randomly selected from a dataset acquired from online sources such as Polyhaven and Laval [11], giving ~4K environment maps that are also flipped and rotated randomly during training. At test time, we employ StableMaterialMV [23], taken off the shelf, to infer materials from the input images to condition the relighting diffusion model. Given that StableMaterialMV was trained on 256×256 images, we separately finetune it further to 512×512 to estimate high quality material maps.

Metrics. For both the 2D and 3D relighting evaluations, we account for the underlying albedo scale ambiguity by a scale against the groundtruth image before computing the PSNR, SSIM, and LPIPS metrics. We also report the approximate time from start to finish LightSwitch and other methods need to predict relightings. Similarly to previous work, the final results are computed as the mean across views for all objects.

4.2. Image to Image Relighting

To validate LightSwitch's relighting and denoising scheme, we evaluate with source RGB images of synthetic test objects captured under an unknown illumination and relight them with a target lighting condition not included in the training environment map dataset. We render the objects under the corresponding target illuminations and compare the relit predictions to the groundtruth appearance. We utilize 6 diverse test objects from the BlenderVault dataset excluded from training and render 8 randomly sampled views on a hemisphere with the object at its center. The object appearances are rendered given 3 fixed illuminations for those 8 views, giving a total of 144 test images.

Baselines. We test LightSwitch against other diffusion-based relighting methods [17, 42] trained to predict relighting for object data from a single image and report the relighting comparison after rescaling. To highlight the relighting consistency across multiple views, in addition to the typical image level rescaling (ILR) metric that searches for optimal rescaling for each image, we also report a stricter scene level rescaling (SLR) metric that computes a single scale across all views in the scene, penalizing inconsistent predictions across views.

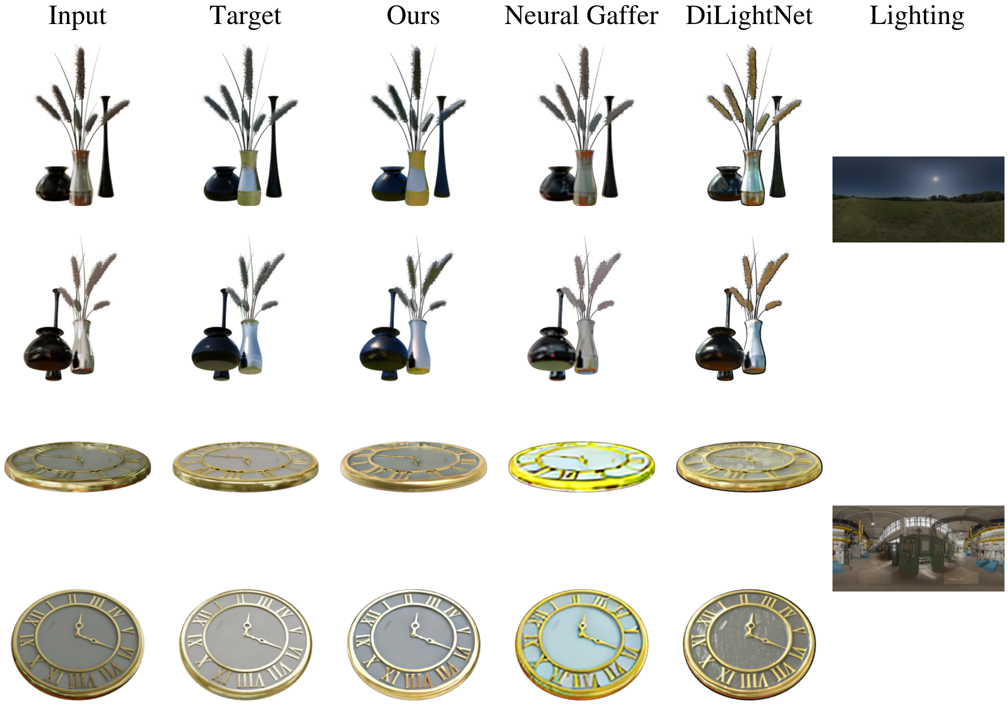

Figure 4. **Direct Relighting Comparison on Synthetic Objects.** Given 8 images of an object, LightSwitch predicts a multi-view consistent relighting under a target illumination. With its usage of inferred material information, our model accurately relights objects with complex appearance effects such as specularities. On the other hand, the baselines bake in details from the source view into the target relighting and relight inconsistently across views.

	Image Relighting (ILR)			**Image Relighting (SLR)**			**Quality Drop**
Method	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↓
DiLightNet [42]	23.84	0.861	0.238	23.35	0.859	0.238	0.49
Neural Gaffer [17]	24.34	0.883	0.271	24.08	0.882	0.272	0.26
Ours	26.01	0.888	0.216	25.86	0.885	0.215	0.15
Ours (GT Materials)	28.29	0.901	0.203	28.20	0.901	0.203	0.09

Table 1. **Direct Relighting Accuracy Comparison.** We report the performance of our approach for 2D relighing on synthetic object data against other diffusion-based relighting methods. ILR corresponds to image level rescaling where the images are individually rescaled against the groundtruth image. SLR corresponds to scene level rescaling where we rescale using the average scale for all views per object. In each column, the best , second best , and third best results are marked.

Results. We report quantitative results in Tab. 1, comparing LightSwitch with previous image relighting methods. Overall, the gains over baselines highlight the benefits of our design choices of leveraging multi-view attention and material priors for relighting objects with complex and diverse materials. In particular, our method achieves the most consistency when using groundtruth materials, indicating its ability to exploit material information for a more consistent and accurate relighting. Fig. 4 shows a qualitative comparison of our model's relightings against baselines.

4.3. Relighting for 3D

We evaluate the novel view relighting quality and efficiency of LightSwitch against prior inverse rendering methods on a number of publicly available diverse synthetic and real object datasets. For synthetic objects, we directly render and compare the groundtruth relightings under novel environment maps, while real objects were captured in novel environments for which the illuminations are estimated.

Datasets. We use the NeRF synthetic dataset (5 objects) [27] and real object dataset Objects with Lighting (8 ob-

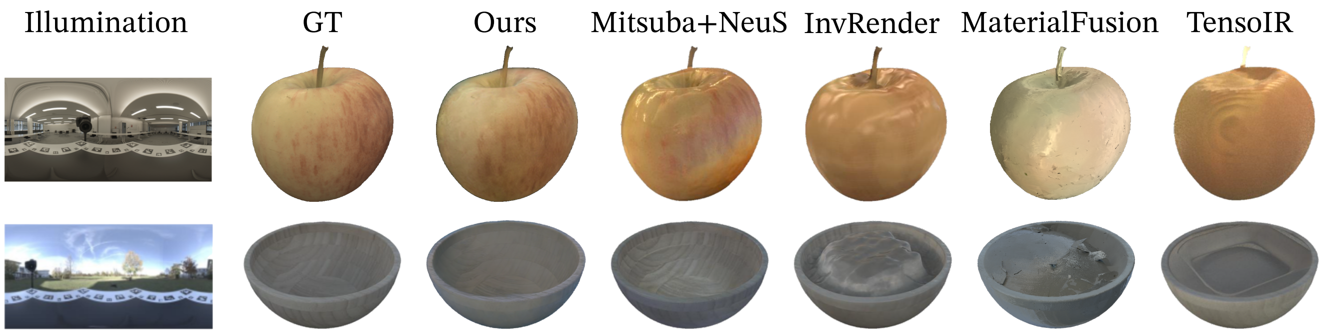

Figure 5. **3D Relighting Comparison on Objects With Lighting.** Our method successfully relights a novel view to a target illumination while the baselines exhibit errors in the relit appearance. LightSwitch's efficiency means it can relight a given novel view in 5 minutes at a high accuracy while operating on images at the original resolution (1728×1120).

	Relighting		
Method	PSNR ↑	SSIM ↑	LPIPS ↓
LightSwitch	25.43	0.84	0.297
Mitsuba+NeuS [15, 38]	26.24	0.84	0.227
InvRender [46]	23.45	0.77	0.374
NeRD [4]	21.71	0.65	0.540
NeRFactor [45]	20.62	0.72	0.486
NeROIC [20]	21.59	0.78	0.323
Neural-PIL [5]	19.56	0.51	0.604
NVDiffrec [29]	22.60	0.72	0.406
NVDiffrecMC [12]	20.24	0.73	0.393
PhySG [43]	22.77	0.82	0.375
TensoIR [16]	24.15	0.77	0.378
MaterialFusion [23]	20.75	0.73	0.388

Table 2. **Relighting on the Objects With Lighting Dataset.** Our method matches and outperforms a multitude of inverse rendering baselines on novel view relighting across the Objects With Lighting dataset.

jects) [34]. NeRF synthetic objects are relit by four high resolution environment maps, and the relighting comparison is computed on a test set of eight unseen poses per environment map. For Objects with Lighting, objects are relit with two novel environment maps for six test views, three per environment map. We show our method's relighting quantitatively and qualitatively on both of these datasets to highlight its strong performance on synthetic and real data.

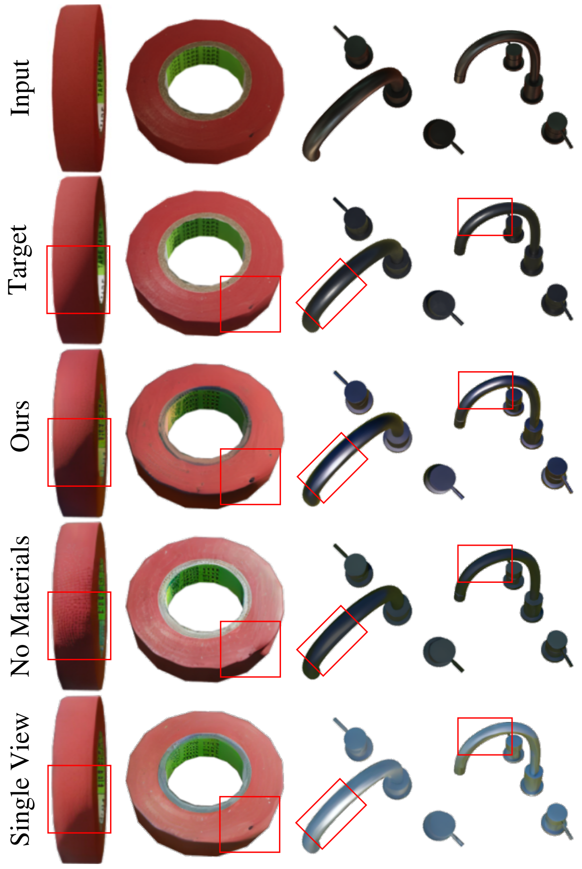

Figure 6. **2D Relighting Ablation Comparison.**

Results. We benchmark LightSwitch against a suite of inverse rendering baselines on both datasets. As shown in Tab. 2 and Tab. 3, our trained model matches or outperforms state-of-the-art methods in relighting across multiple synthetic and real objects. Our model can successfully relight objects exhibiting complex appearance properties. By distributing denoising across compute with our denoising scheme, we relight in much less time than the next best performing baseline, which takes orders of magnitude longer than ours to do relighting. Fig. 7 and Fig. 5 show a qualitative comparison of our model against previous approaches on both datasets. We compare the runtime for our method using 8 RTX A6000 and the other methods that can only utilize 1 RTX A6000 and find that our method is proportionally faster while outperforming or matching baselines.

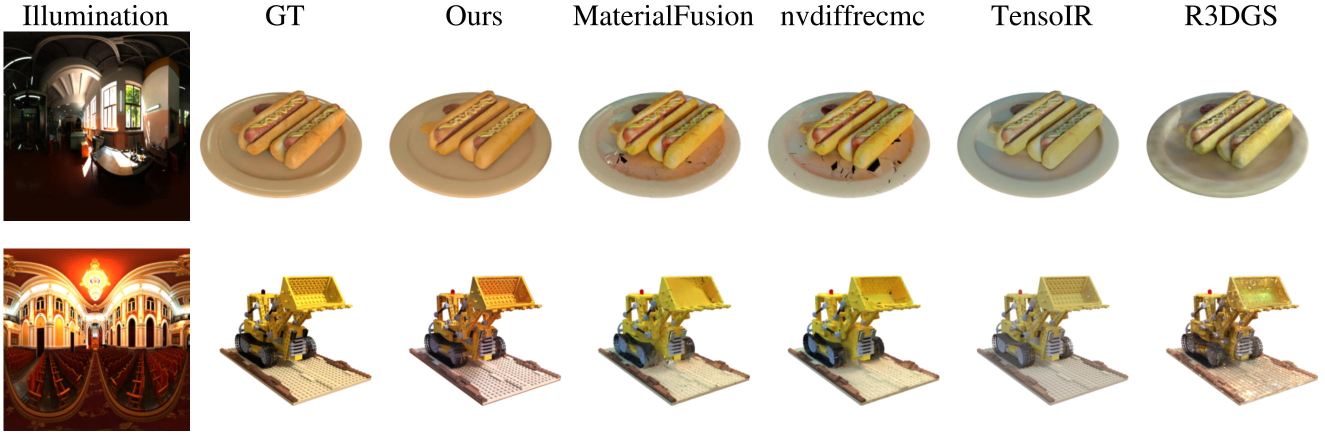

Figure 7. **3D Relighting Comparison on NeRF-Synthetic.** While other methods exhibit issues in the relit appearance such as baked-in albedo, reconstruction artifacts, and incorrect geometry, our method successfully relights with high fidelity.

	Chair		**Hotdog**		**Lego**		**Materials**		**Mic**		**Runtime**
Method	PSNR	LPIPS	PSNR	LPIPS	PSNR	LPIPS	PSNR	LPIPS	PSNR	LPIPS	Minutes
LightSwitch	26.65	0.062	25.75	0.091	23.60	0.081	22.08	0.080	30.24	0.025	~2
MaterialFusion [23]	26.58	0.063	25.31	0.123	23.26	0.119	25.29	0.084	30.94	0.036	~240
NVDiffrecMC [12]	26.44	0.064	24.87	0.133	23.36	0.115	25.37	0.081	30.15	0.041	~120
TensoIR [16]	25.29	0.070	21.16	0.174	21.86	0.080	22.02	0.104	31.21	0.022	~480
R3DGS [10]	23.50	0.072	21.02	0.168	20.86	0.106	20.56	0.095	29.47	0.029	~15

Table 3. **Relighting on the NeRF-Synthetic Dataset.** Our method matches or outperforms the baselines on novel view relighting across all views per NeRF-Synthetic object at a much lower runtime.

Ablations	PSNR ↑	SSIM ↑	LPIPS ↓
Ours	26.01	0.888	0.216
No Materials	25.27	0.879	0.219
Single View	24.59	0.865	0.228

Table 4. **Effects of Ablating Multi-view or Materials.** Ablating different information from the relighting diffusion framework harms relighting accuracy.

Using 1 GPU increases runtime to 14 minutes without a loss in quality, which is comparable to R3DGS [10] in runtime but produces much more accurate relighting. The strong performance in both 2D and 3D relighting exhibited by our method showcases its utilization of multi-view appearance and material cues for effective relighting.

4.4. Ablation Studies

We finetune additional diffusion models that ablate multi-view or material information incorporation into the architecture and evaluate on the BlenderVault 2D test dataset. The quantitative and qualitative comparisons are shown in Tab. 4 and Fig. 6. Not giving material information during training leads to a considerable drop in quality, as the model struggles with relighting more complex appearances such as specularities, and begins incorporating details from the input views to the predicted relightings. Training with materials but not with multi-view leads to an even bigger drop, as the model struggles to produce consistent relightings.

5. Conclusion

In this paper, we introduced LightSwitch, a generative relighting framework capable of leveraging inferred material cues for accurate and consistent multi-view relighting. While this improved over prior works in 2D and 3D relighting, we believe there are some remaining limitations. First, as shown in the appendix, the reliance on the (fixed) latent space of a pre-trained diffusion limits the ability to encode/decode sharp fine details *e.g.* reflections. Moreover, while our architecture encourages multi-view consistency and material-aware inference, the predictions are not guaranteed to be physically plausible. We believe that exploring alternate architectures and mechanisms to more closely connect learned relighting with physics-based rendering are promising avenues for future exploration.

6. Acknowledgments

This work was supported in part by the NSF GFRP (Grant No. DGE2140739) and NSF Award IIS-2345610. This work used Bridges-2 at Pittsburgh Supercomputing Center through allocation CIS240022 from the ACCESS program, which is supported by National Science Foundation grants #2138259, #2138286, #2138307, #2137603, and #2138296.

References

[1] Hadi Alzayer, Philipp Henzler, Jonathan T. Barron, Jia-Bin Huang, Pratul P. Srinivasan, and Dor Verbin. Generative multiview relighting for 3d reconstruction under extreme illumination variation. In *CVPR*, 2025. 2

[2] Jonathan T. Barron, Ben Mildenhall, Dor Verbin, Pratul P. Srinivasan, and Peter Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In *CVPR*, 2022. 2

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

[4] Mark Boss, Raphael Braun, Varun Jampani, Jonathan T. Barron, Ce Liu, and Hendrik P.A. Lensch. Nerd: Neural reflectance decomposition from image collections. In *ICCV*, 2021. 7

[5] 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. In *NeurIPS*, 2021. 2, 7

[6] Brent Burley and Walt Disney Animation Studios. Physically-based shading at disney. In *SIGGRAPH*, 2012. 3

[7] Yuedong Chen, Chuanxia Zheng, Haofei Xu, Bohan Zhuang, Andrea Vedaldi, Tat-Jen Cham, and Jianfei Cai. Mvsplat360: Feed-forward 360 scene synthesis from sparse views. In *NeurIPS*, 2024. 2

[8] 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. In *CVPR*, 2023. 5

[9] Ye Fang, Zeyi Sun, Shangzhan Zhang, Tong Wu, Yinghao Xu, Pan Zhang, Jiaqi Wang, Gordon Wetzstein, and Dahua Lin. Relightvid: Temporal-consistent diffusion model for video relighting. *arXiv*, 2025. 2

[10] Jian Gao, Chun Gu, Youtian Lin, Hao Zhu, Xun Cao, Li Zhang, and Yao Yao. Relightable 3d gaussian: Real-time point cloud relighting with brdf decomposition and ray tracing. In *ECCV*, 2024. 2, 8

[11] Marc-André Gardner, Kalyan Sunkavalli, Ersin Yumer, Xiaohui Shen, Emiliano Gambaretto, Christian Gagné, and Jean-François Lalonde. Learning to predict indoor illumination from a single image. *ACM Transactions on Graphics (ToG)*, 2017. 5

[12] Jon Hasselgren, Nikolai Hofmann, and Jacob Munkberg. Shape, Light, and Material Decomposition from Images using Monte Carlo Rendering and Denoising. In *NeurIPS*, 2022. 2, 7, 8

[13] Binbin Huang, Zehao Yu, Anpei Chen, Andreas Geiger, and Shenghua Gao. 2d gaussian splatting for geometrically accurate radiance fields. In *SIGGRAPH*, 2024. 2

[14] Xin Huang, Tengfei Wang, Ziwei Liu, and Qing Wang. Material anything: Generating materials for any 3d object via diffusion. In *CVPR*, 2025. 5

[15] Wenzel Jakob, Sébastien Speierer, Nicolas Roussel, Merlin Nimier-David, Delio Vicini, Tizian Zeltner, Baptiste Nicolet, Miguel Crespo, Vincent Leroy, and Ziyi Zhang. Mitsuba 3 renderer, 2022. https://mitsuba-renderer.org. 7

[16] Haian Jin, Isabella Liu, Peijia Xu, Xiaoshuai Zhang, Songfang Han, Sai Bi, Xiaowei Zhou, Zexiang Xu, and Hao Su. Tensoir: Tensorial inverse rendering. In *CVPR*, 2023. 2, 7, 8

[17] 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. In *NeurIPS*, 2024. 2, 3, 5, 6

[18] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics (ToG)*, 2023. 1, 2, 5

[19] Peter Kocsis, Vincent Sitzmann, and Matthias Nießner. Intrinsic image diffusion for indoor single-view material estimation. In *CVPR*, 2024. 2

[20] Zhengfei Kuang, Kyle Olszewski, Menglei Chai, Zeng Huang, Panos Achlioptas, and Sergey Tulyakov. Neroic: Neural rendering of objects from online image collections. *ACM Transactions on Graphics (ToG)*, 2022. 7

[21] Ruofan Liang, Zan Gojcic, Huan Ling, Jacob Munkberg, Jon Hasselgren, Zhi-Hao Lin, Jun Gao, Alexander Keller, Nandita Vijaykumar, Sanja Fidler, and Zian Wang. Diffusionrenderer: Neural inverse and forward rendering with video diffusion models. *CVPR*, 2025. 2

[22] Zhihao Liang, Qi Zhang, Ying Feng, Ying Shan, and Kui Jia. Gs-ir: 3d gaussian splatting for inverse rendering. In *CVPR*, 2024. 2

[23] Yehonathan Litman, Or Patashnik, Kangle Deng, Aviral Agrawal, Rushikesh Zawar, Fernando De la Torre, and Shubham Tulsiani. Materialfusion: Enhancing inverse rendering with material diffusion priors. In *3DV*, 2025. 2, 3, 5, 7, 8

[24] Minghua Liu, Ruoxi Shi, Linghao Chen, Zhuoyang Zhang, Chao Xu, Xinyue Wei, Hansheng Chen, Chong Zeng, Jiayuan Gu, and Hao Su. One-2-3-45++: Fast single image to 3d objects with consistent multi-view generation and 3d diffusion. In *CVPR*, 2024. 2

[25] Xiaoxiao Long, Yuan-Chen Guo, Cheng Lin, Yuan Liu, Zhiyang Dou, Lingjie Liu, Yuexin Ma, Song-Hai Zhang, Marc Habermann, Christian Theobalt, et al. Wonder3d: Single image to 3d using cross-domain diffusion. In *CVPR*, 2024. 2

[26] Xiaoyang Lyu, Yang-Tian Sun, Yi-Hua Huang, Xiuzhe Wu, Ziyi Yang, Yilun Chen, Jiangmiao Pang, and Xiaojuan Qi. 3dgsr: Implicit surface reconstruction with 3d gaussian splatting. *ACM Transactions on Graphics (TOG)*, 2024. 2

[27] 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. In *ECCV*, 2020. 1, 2, 6

[28] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Transactions on Graphics (ToG)*, 2022. 2

[29] Jacob Munkberg, Jon Hasselgren, Tianchang Shen, Jun Gao, Wenzheng Chen, Alex Evans, Thomas Müller, and Sanja Fidler. Extracting Triangular 3D Models, Materials, and Lighting From Images. In *CVPR*, 2022. 2, 7

[30] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 3

[31] Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. In *ICLR*, 2021. 4

[32] Yichun Shi, Peng Wang, Jianglong Ye, Mai Long, Kejie Li, and Xiao Yang. Mvdream: Multi-view diffusion for 3d generation. In *ICLR*, 2024. 4

[33] 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. In *CVPR*, 2021. 2

[34] Benjamin Ummenhofer, Sanskar Agrawal, Rene Sepúlveda, Yixing Lao, Kai Zhang, Tianhang Cheng, Stephan R. Richter, Shenlong Wang, and Germán Ros. Objects with lighting: A real-world dataset for evaluating reconstruction and rendering for object relighting. In *3DV*, 2024. 7

[35] 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. In *SIGGRAPH Asia*, 2024. 2

[36] Vikram Voleti, Chun-Han Yao, Mark Boss, Adam Letts, David Pankratz, Dmitry Tochilkin, Christian Laforte, Robin Rombach, and Varun Jampani. Sv3d: Novel multi-view synthesis and 3d generation from a single image using latent video diffusion. In *ECCV*, 2024. 2

[37] Peng Wang and Yichun Shi. Imagedream: Image-prompt multi-view diffusion for 3d generation. *arXiv*, 2023. 4

[38] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. In *NeurIPS*, 2021. 7

[39] Rundi Wu, Ben Mildenhall, Philipp Henzler, Keunhong Park, Ruiqi Gao, Daniel Watson, Pratul P Srinivasan, Dor Verbin, Jonathan T Barron, Ben Poole, et al. Reconfusion: 3d reconstruction with diffusion priors. In *CVPR*, 2024. 2

[40] Chen Xi, Peng Sida, Yang Dongchen, Liu Yuan, Pan Bowen, Lv Chengfei, and Zhou. Xiaowei. Intrinsicanything: Learning diffusion priors for inverse rendering under unknown illumination. In *ECCV*, 2024. 2

[41] Zhiwen Yan, Weng Fei Low, Yu Chen, and Gim Hee Lee. Multi-scale 3d gaussian splatting for anti-aliased rendering. In *CVPR*, 2024. 2

[42] Chong Zeng, Yue Dong, Pieter Peers, Youkang Kong, Hongzhi Wu, and Xin Tong. Dilightnet: Fine-grained lighting control for diffusion-based image generation. In *SIGGRAPH*, 2024. 2, 5, 6

[43] Kai Zhang, Fujun Luan, Qianqian Wang, Kavita Bala, and Noah Snavely. PhySG: Inverse rendering with spherical gaussians for physics-based material editing and relighting. In *CVPR*, 2021. 7

[44] Tianyuan Zhang, Zhengfei Kuang, Haian Jin, Zexiang Xu, Sai Bi, Hao Tan, He Zhang, Yiwei Hu, Milos Hasan, William T Freeman, et al. Relitlrm: Generative relightable radiance for large reconstruction models. In *ICLR*, 2025. 2

[45] 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. *ACM Transactions on Graphics (ToG)*, 2021. 2, 7

[46] Yuanqing Zhang, Jiaming Sun, Xingyi He, Huan Fu, Rongfei Jia, and Xiaowei Zhou. Modeling indirect illumination for inverse rendering. In *CVPR*, 2022. 2, 7

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

[48] Liu Zhenyuan, Yu Guo, Xinyuan Li, Bernd Bickel, and Ran Zhang. Bigs: Bidirectional gaussian primitives for relightable 3d gaussian splatting. In *3DV*, 2025. 2

[49] Zhizhuo Zhou and Shubham Tulsiani. Sparsefusion: Distilling view-conditioned diffusion for 3d reconstruction. In *CVPR*, 2023. 2