MaterialFusion: Enhancing Inverse Rendering with Material Diffusion Priors

Yehonathan Litman[1] Or Patashnik[2] Kangle Deng[1] Aviral Agrawal[1]
Rushikesh Zawar[1] Fernando De la Torre[1] Shubham Tulsiani[1]
[1]Carnegie Mellon University [2]Tel Aviv University

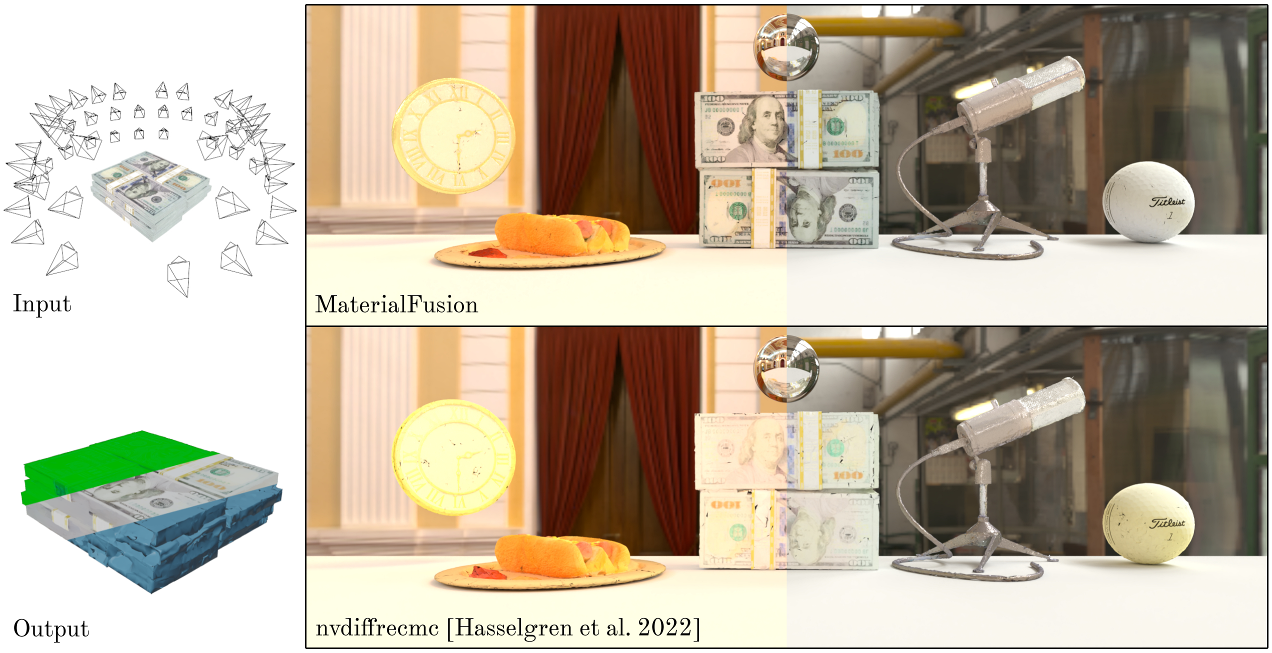


Figure 1. Given an image set of an object under unknown illumination, MaterialFusion recovers the object's geometry, BRDF appearance, and the environmental illumination, via inverse rendering. Our method utilizes a 2D material diffusion prior to accurately reconstruct these properties. On the left, we display the input image set of the bills alongside the output of the reconstructed properties, visualized as the materials, albedo, and mesh from top to bottom, respectively. On the right, we show different objects rendered under novel lighting conditions with the reconstructed physical properties.

Abstract

Recent works in inverse rendering have shown promise in using multi-view images of an object to recover shape, albedo, and materials. However, the recovered components often fail to render accurately under new lighting conditions due to the intrinsic challenge of disentangling albedo and material properties from input images. To address this challenge, we introduce MaterialFusion, an enhanced conventional 3D inverse rendering pipeline that incorporates a 2D prior on texture and material properties. We present StableMaterial, a 2D diffusion model prior that refines multi-lit data to estimate the most likely albedo and material from given input appearances. This model is trained on albedo, material, and relit image data derived from a curated dataset of approximately ~12K artist-designed synthetic Blender objects called BlenderVault. We incorporate this diffusion prior with an inverse rendering framework where we use score distillation sampling (SDS) to guide the optimization of the albedo and materials, improving relighting performance in comparison with previous work. We validate MaterialFusion's relighting performance on 4 datasets of synthetic and real objects under diverse illumination conditions, showing our diffusion-aided approach significantly improves the appearance of reconstructed objects under novel lighting conditions. We intend to publicly release our BlenderVault dataset to support further research in this field.

1. Introduction

Recently, there has been an increased interest in methods that try to recover 3D representations from 2D images. Novel view synthesis approaches, particularly Neural Radiance Fields (NeRF) [32] and follow-up works have proven highly effective for accurately representing 3D scenes from posed 2D images. Nevertheless, one of the main drawbacks of these approaches is relighting, since novel view synthesis methods bake in all the lighting information into the 3D representation, rather than disentangling it from the underlying scene data. In this paper, our goal is to infer relightable 3D representations that factorize these properties, allowing for the editing of materials, geometry, and lighting independently.

Some approaches do pursue factorized, relightable 3D representations [5, 17, 53]. These methods employ signed distance functions (SDFs), meshes, or volumetric representations to model geometry, while also estimating underlying properties like diffuse albedo and specular parameters and their results can be used for relighting in novel environments. However, as these approaches are supervised on captured image data under a fixed illumination, there still exists an ambiguity between the underlying properties that images alone cannot explain. Multiple possible materials and textures could be composed onto the geometry to produce the same final images in the training data, leading to fundamental ambiguities when inferring underlying albedo and material properties using a simple pixel-level reconstruction loss. The ill-posed nature of this problem ultimately leads to suboptimal factorization.

Our key insight is that 2D priors over plausible materials and albedos, in addition to reconstruction losses, can resolve ambiguities in factorized inverse rendering. We learn a large scale conditional diffusion prior over likely materials for RGB images under different illuminations. In addition to reconstruction loss, we distill the fine-tuned diffusion model to provide additional signal about plausible texture and material combinations for the depicted object during 3D optimization.

We demonstrate our 3D inverse rendering approach, MaterialFusion, on the NeRF Synthetic, NeRFactor datasets [32, 58], a test set of our BlenderVault dataset and the Stanford-ORB dataset [23]. We use these datasets to show significant improvements in novel view synthesis under relighting as well as material estimation compared to prior state-of-the-art work on both synthetic and real data. We trained a conditional diffusion model, StableMaterial, with albedos, materials, and relit images rendered from ~29K high quality objects, augmented with our own BlenderVault dataset of ~12K high quality synthetic Blender objects curated from online sources, and show its superior performance compared to previous approaches that recover albedo and materials from input images. Using our prior, our learned factorized representation generalizes better to novel lighting conditions across diverse lighting, object, and underlying material scenarios, as shown in the relighting results in Fig. 1.

2. Related Works

2.1. Inverse Rendering

In recent years, reconstruction methods that learn a 3D representation from a set of multi-view images have rapidly improved [1, 2, 21, 32, 34] in terms of quality and speed. However, many of these methods do not disentangle the underlying texture and materials, from the illumination. Therefore, rendering the acquired scene under novel lighting conditions remains a challenge.

To address this, inverse rendering works have begun focusing on reconstructing the 3D appearance along with the underlying materials of a scene or object. Given a set of images of a scene or object under a fixed illumination, some works have aimed to recover the texture, materials, and lighting [4, 14, 17, 20, 24, 35, 58, 59]. This task is inherently challenging due to its high dimensionality and ambiguity in explaining the image appearance, as multiple illumination and material parameter combinations can be used to reproduce the final appearance. To tackle this ambiguity, other works simplify the problem setting by assuming or modeling scene lighting [10, 16, 20, 57] or employing domain-specific priors [3, 9, 58] to inject additional information on physical properties. Nevertheless, assumptions about lighting limit the applicability of these methods in real-world scenarios such as online marketplaces, where lighting conditions can be difficult to capture and are constantly changing. Moreover, the priors used in the aforementioned works are either trained on small-scale or procedurally generated data or focus on a specific object category.

In contrast, our approach does not rely on controlled lighting conditions; instead, it primarily utilizes a large-scale 2D texture and material prior trained with a large synthetic object dataset we curated. The objects in this dataset contain complex Physically Based Rendering (PBR) assets, enhancing our prior's predictions.

2.2. 2D Diffusion Priors For 3D Tasks

The success of diffusion models in text-to-image synthesis [18, 39, 40] has also brought attention to employing large scale 2D priors for 3D generation [8, 25, 31, 37, 46, 49, 51]. Dreamfusion [37] and SJC [49] first propose Score Distillation Sampling (SDS) to optimize a 3D representation using 2D diffusion model gradients. Some follow-up works enriched the 2D model prior with 3D knowledge by fine-tuning the model to generate novel views of an object [28], to generate images of several views simultaneously [29, 45, 50]. Moreover, it has been shown that such enriched models perform better in generating 3D models from scratch and in single-view reconstruction [26–30, 44, 45, 50, 60]. Additionally, ReconFusion [54] also uses 2D diffusion priors to improve sparse-view 3D reconstruction. However, common to all of these works is the lack of material and illumination disentanglement, thereby limiting the relighting performance of the generated or reconstructed objects.

To predict physical properties, previous works showed success in finetuning a pretrained diffusion model. Specifically, some works predict material parameters given an RGB image [22, 42, 48, 56]. However, these works reconstruct only a 2D representation of the underlying physical properties, and do not consider the 3D reconstruction from a set of images. In contrast to the aforementioned works, our approach reconstructs the underlying 3D geometry, material properties, and environmental lighting from a set of multi-view images via score distillation. Closest to ours, [55] concurrently used a 2D diffusion model to guide relightable 3D inference, but used diffusion samples to guide the optimization while ours uses likelihood maximization via SDS.

2.3. 3D Datasets with PBR assets

The availability of 3D datasets is considerably smaller than the availability of 2D datasets, even more so in terms of PBR information, imposing a challenge in 3D-related tasks. In particular, commonly used 3D datasets [13, 19, 38] lack PBR information. Some datasets [11, 36] offer 3D objects with PBR information but are limited in diversity to only furniture. Objaverse [12] offers diversity yet contains many objects that are partial reconstructions, low in quality, or cartoonish. Artist-designed high-quality 3D objects with PBR data are available in different sources, but are not organized in a dataset suitable for research. In this work, we introduce a new dataset of Blender objects containing high quality PBR assets curated from online sources. We use this dataset to augment previous datasets, greatly enhancing the diversity of PBR information available for training.

3. Methodology

This section introduces MaterialFusion, our approach to reconstructing a 3D representation of an object from a set of multi-view images. An overview of our approach is shown in Fig. 3. Specifically, given a set of posed images of the object captured under an unknown illumination, our goal is to reconstruct the object's geometry and BRDF appearance, as well as recover the environmental illumination. Accurately reconstructing these components allows us to faithfully recreate the object appearance under new lighting conditions. We represent the geometry as a mesh, as its explicit nature is more suitable for downstream tasks. For the material, we use a simplified Disney principled BRDF model [7] representation. Specifically, the material texture contains three components per texel, albedo $\mathbf{a} \in \mathbb{R}^3$, roughness $r \in \mathbb{R}$, and metallicness $m \in \mathbb{R}$. Following prior works [17, 35], we represent the albedo texture as an albedo UV-map $\mathbf{k}_\text{d}$, and roughness and metallicness as part of an occlusion, roughness, metallicness (ORM) UV-map $\mathbf{k}_\text{orm}$, where each texel is (o, r, m) with o unused. The environment illumination is represented as a high dynamic range (HDR) environment map.

Our key idea is to leverage a strong 2D prior obtained from an image diffusion model which is trained to estimate the underlying material given a RGB image input. To accomplish this, we first adapt an existing image diffusion model (Stable Diffusion 2.1 [39]) to predict the albedo and ORM from an image of an arbitrary object rendered under a randomly selected illumination. This allows us to extend an existing 2D diffusion prior such that it has material understanding. The finetuning procedure is shown in Fig. 2. We then leverage the extended 2D prior in an inverse rendering framework to infer a disentangled 3D representation of a given object and an HDR map of the environment lighting. Specifically, we utilize a variant of SDS loss [37] to employ the 2D prior for 3D optimization. We show an overview of the 3D inference procedure in Fig. 3.

3.1. Training Data

Learning a diffusion prior for albedo and ORM prediction from images we leverage a diverse dataset of synthetic object renderings with high-quality PBR textures. Using such data, we generate training images with a graphics engine capable of reproducing realistic appearances such as Blender. We examined existing datasets such as Objaverse [12] and ABO [11] for this purpose.

Objaverse is a large and diverse dataset, but it contains many unrealistic, low-quality, or textureless objects. To address this, we followed a similar filtering procedure as [47], and then further filtered for non-cartoon objects with PBR textures. This resulted in a filtered subset of ~8.5K objects from Objaverse.

While the filtered Objaverse dataset provided good coverage, we found that augmenting it with the ABO dataset (which contains ~8K objects from only 63 categories) was not sufficient to achieve the desired diversity in our training data. To further improve the diversity, we created our own BlenderVault dataset, which contains an additional ~12K high-quality, PBR-textured objects. BlenderVault consists of Blender objects designed and validated by artists across arbitrary categories for use in commercial projects.

To render the training images, we replaced any glass surfaces in the objects with a black surface of roughness 0.25 and metallicness 0. We then rendered 30 images of each object, with randomly selected azimuth $\sim [0°, 360°]$ and elevation $\sim [-15°, 90°]$ on a hemisphere with a radius $\sim [1.5, 2.0]$. The lighting conditions were also varied,

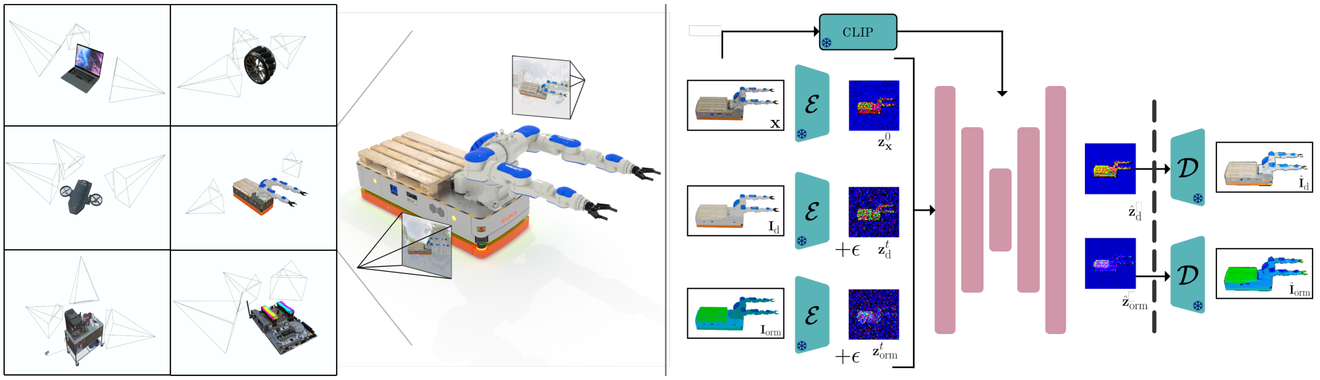


Figure 2. StableMaterial receives an RGB image as input and outputs the albedo $\hat{\mathbf{I}}_\text{d}$ and ORM $\hat{\mathbf{I}}_\text{orm}$ 2D maps. To train StableMaterial, we use BlenderVault objects to render a dataset of multi-view images under varying illuminations as well as the corresponding albedo and ORM maps. Given a triplet $(\mathbf{x}, \mathbf{I}_\text{d}, \mathbf{I}_\text{orm})$ of an image and its albedo and ORM maps, we encode them using the pretrained Stable Diffusion encoder and concatenate the image latent with the noisy albedo and ORM latents. The model is then trained with a diffusion loss to denoise the albedo and ORM maps.

using a random selection of StreetLearn [33] environment maps, Laval [15] indoor environment maps, point lights, or directional sun lights.

In total, our training dataset consists of ~28K synthetic objects with high quality PBR assets, combining the filtered Objaverse, ABO, and our own BlenderVault data.

3.2. StableMaterial – 2D Material Denoising Diffusion Prior

To have a strong 2D material prior, we build on Stable Diffusion 2.1 [39] and fine-tune it from the pretrained model on a dataset consisting of triplets $(\mathbf{x}, \mathbf{I}_\text{d}, \mathbf{I}_\text{orm})$, where $\mathbf{x}$ is an RGB image of the object, and $\mathbf{I}_\text{d}, \mathbf{I}_\text{orm}$ are its corresponding rendered albedo and ORM components, respectively. Formally, given an RGB input $\mathbf{x}$ of an object under unknown illumination, we fine-tune Stable Diffusion to output its underlying albedo $\mathbf{I}_\text{d}$ and ORM $\mathbf{I}_\text{orm}$ components.

Model Architecture. We modify only Stable Diffusion's UNet, so that it is conditioned on the input image $\mathbf{x}$ in two ways. First, we encode it with Stable Diffusion's pre-trained frozen VAE $\mathcal{E}$ and concatenate the resulting clean latent $\mathbf{z}_\mathbf{x}^0$ to the noisy latent codes $\mathbf{z}^t$ in the channel dimension, where t is the diffusion timestep. Specifically, the noisy latent code $\mathbf{z}^t = [\mathbf{z}_\text{d}^t, \mathbf{z}_\text{orm}^t]$, i.e. the concatenation of the noisy albedo latent $\mathbf{z}_\text{d}^t$ and the noisy ORM map $\mathbf{z}_\text{orm}^t$ in the channel dimension. The input of our UNet is then $\left(\mathbf{z}_\mathbf{x}^0, \mathbf{z}^t, t\right)$. The text conditioning is also replaced with a CLIP image embedding of the input image. These two ways of inputting the image into the model allow it to have both global and local reasoning about the image.

To output both albedo and ORM maps, our noisy latent codes $\mathbf{z}^t$ consist of 8 channels, 4 corresponding to the albedo and the other 4 corresponding to the ORM. In total, the input of our network is composed of 12 channels consisting of the encoded input image, noisy albedo latents and the noisy ORM latents. To obtain the RGB albedo and ORM maps we decode the denoised $\hat{\mathbf{z}}$ through the pretrained Stable Diffusion decoder $\mathcal{D}$. To account for the different input and output channels, the first and last layers are changed and randomly initialized, while the other UNet parameters are kept unchanged.

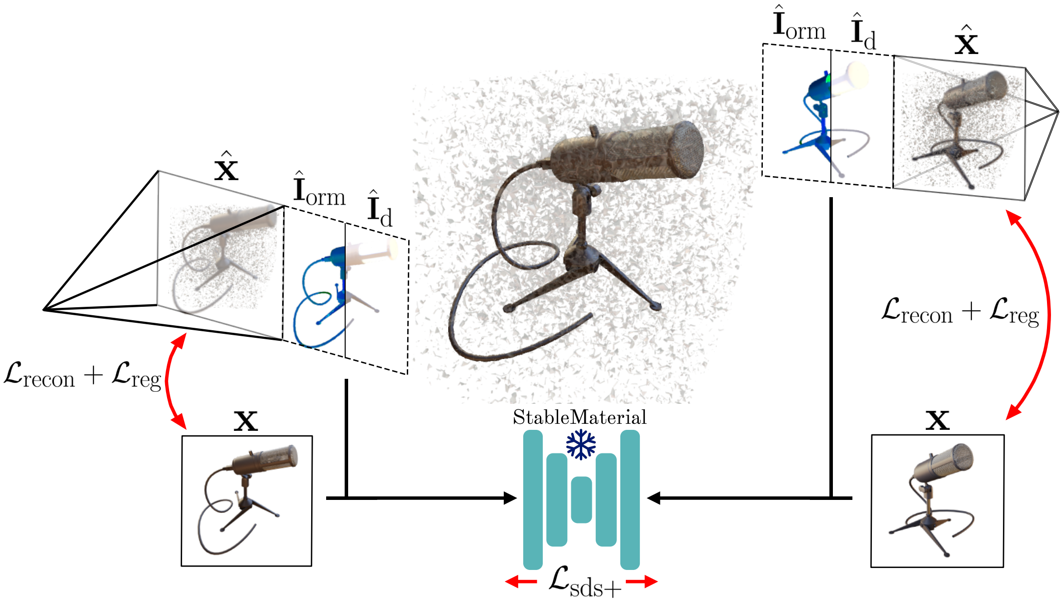
Loss. To fine-tune the model, we utilize v-prediction diffusion loss [41]. At each training iteration, we sample a triplet $(\mathbf{x}, \mathbf{I}_\text{d}, \mathbf{I}_\text{orm})$, and encode each of the images with $\mathcal{E}$. We concatenate $\mathcal{E}(\mathbf{I}_\text{d}), \mathcal{E}(\mathbf{I}_\text{orm})$ in the channel dimension and denote their concatenation by $\mathbf{z}$. We sample a diffusion timestep t along with an 8-channel random noise ϵ, and add the noise to $\mathbf{z}$ to obtain $\mathbf{z}^t$. The diffusion loss is defined as

$$\mathcal{L}_\text{diff} = \mathbb{E}_{\mathbf{x},\mathbf{k}_\text{d},\mathbf{k}_\text{orm},\epsilon\sim\mathcal{N}(0,I),t}\left[\left\|\epsilon_\theta\left(\mathcal{E}(\mathbf{x}), \mathbf{z}^t, t\right) - \mathbf{v}_t\right\|_2^2\right], \tag{1}$$

where $\mathbf{x}, \mathbf{z}^t$ are as defined above. As in [41], $\mathbf{v}_t = \alpha_t\epsilon - \sigma_t\mathbf{z}$, where α_t, σ_t are the parameters of the scheduler. Similarly to [6], we enable classifier-free guidance by setting the input images, input image prompt, or both to all zeros with a 5% probability each and set the guidance scale to 3.0.

3.3. Prior-guided Inverse Rendering

Having material knowledge in our trained StableMaterial model, we can distill this knowledge and reconstruct a 3D disentangled representation of the object. Specifically, given a set of multi-view images under an unknown illumination depicting an object, we aim to reconstruct the underlying geometry represented as a mesh and denoted by $\mathbf{G}$, the underlying UV material texture denoted by $(\mathbf{k}_\text{d}, \mathbf{k}_\text{orm})$, and the environment illumination $\mathbf{L}$. To this end, we directly optimize these representations and build on recent advancements in the distillation of 3D information from 2D



$\hat{\mathbf{x}}$
$\hat{\mathbf{I}}_{\text{orm}}$
$\hat{\mathbf{I}}_{\text{d}}$
$\mathcal{L}_{\text{recon}} + \mathcal{L}_{\text{reg}}$
$\mathbf{x}$
StableMaterial
$\mathcal{L}_{\text{sds+}}$
$\hat{\mathbf{I}}_{\text{orm}}$
$\hat{\mathbf{I}}_{\text{d}}$
$\hat{\mathbf{x}}$
$\mathcal{L}_{\text{recon}} + \mathcal{L}_{\text{reg}}$
$\mathbf{x}$

Figure 3. MaterialFusion reconstructs an object's geometry, PBR materials, and environmental illumination from a set of multi-view images under a fixed lighting condition. In addition to the reconstruction and regularization losses computed between our rendered images $\hat{\mathbf{x}}$ and reference RGB images $\mathbf{x}$, MaterialFusion employs priors from our pre-trained StableMaterial to enhance PBR material reconstruction. Specifically, it calculates an SDS loss for the rendered albedo and ORM components, $\hat{\mathbf{I}}_{\text{d}}$ and $\hat{\mathbf{I}}_{\text{orm}}$ conditioned on $\mathbf{x}$.

diffusion models [37], in conjunction with the off the shelf nvdiffrecmc inverse rendering pipeline [17] as part of MaterialFusion.

Following previous works [17, 35] we parameterize the geometry through an SDF denoted by $\mathbf{S}$, and extract a mesh $\mathbf{G}$ in each optimization iteration using DMTet [43]. Given the mesh $\mathbf{G}$, the texture $(\mathbf{k}_{\text{d}}, \mathbf{k}_{\text{orm}})$, a camera view $\mathbf{C}$, and an HDR environment light map $\mathbf{L}$, we use nvdiffrecmc's differentiable renderer to produce a 2D rendering.

In each optimization iteration, we sample some images and their associated views $\mathbf{x}, \mathcal{C}$, respectively, from the training set, and differentiably render the object using the optimized parameters from the views $\mathcal{C}$. We obtain the rendered image $\hat{\mathbf{x}}$ with nvdiffrecmc's renderer and apply a reconstruction loss to optimize $(\mathbf{S}, \mathbf{k}_{\text{d}}, \mathbf{k}_{\text{orm}}, \mathbf{L})$:

$$\mathcal{L}_{\text{recon}} = \mathbb{E}_{\mathbf{C}}\left[\mathcal{L}_2(\hat{\mathbf{x}}, \mathbf{x})\right], \tag{2}$$

At the same time, we also render the corresponding components of the albedo and the material, $\hat{\mathbf{I}}_{\text{d}}$ and $\hat{\mathbf{I}}_{\text{orm}}$, respectively, encode them with the Stable Diffusion model encoder and concatenate their latent representations to get $\mathbf{z}$. Then, we sample a Gaussian noise ϵ and add it to $\mathbf{z}$ according to a random diffusion timestep $t \in [0.02, 0.98]$, to obtain $\mathbf{z}^t$. We then denoise $\mathbf{z}^t$ using the diffusion model, and sample a clean latent representation of the materials denoted by $\hat{\mathbf{z}}$ via DDIM sampling for 5 steps, conditioned on $\mathbf{x}$. We use the SDS loss in both latent and pixel space:

$$\mathcal{L}_{\text{SDS+}} = \mathbb{E}_{t,\epsilon,v}\left[\lambda_{\text{latent}}||\mathbf{z} - \hat{\mathbf{z}}||^2 + \lambda_{\text{rgb}}||\mathcal{D}(\mathbf{z}) - \mathcal{D}(\hat{\mathbf{z}})||^2\right], \tag{3}$$

where $\mathcal{D}$ is the Stable Diffusion decoder. This loss is inspired by HiFA [61], as we empirically found the RGB term important for boosting the quality of materials estimated during training, shown in Tab. 4. Using 5 denoising steps was key for producing crisp and accurate predictions for the albedo and ORM materials. Finally, our total loss consists of the reconstruction loss, regularization loss, and SDS loss:

$$\mathcal{L}_{\text{MaterialFusion}} = \mathcal{L}_{\text{recon}} + \mathcal{L}_{\text{reg}} + \gamma_i \mathcal{L}_{\text{SDS+}}, \tag{4}$$

where γ_i is an iteration dependent hyperparameter which decays as training progresses. We find that reducing the weight on SDS loss towards the end of the optimization helps preserve finer details that may be lost due to the encoder/decoder operation. We use the same $\mathcal{L}_{\text{reg}}$ as nvdiffrecmc [17]. Conceptually, using SDS loss during inverse rendering maximizes the likelihood of the ORM and albedo under our prior for all training images.

4. Experiments

We evaluate MaterialFusion and StableMaterial on image sequences of objects made of various materials and textures and show the qualitative and quantitative comparisons. We first evaluate MaterialFusion against prior inverse rendering methods for object relighting on a number of synthetic and real diverse objects and highlight the advantages of our approach in terms of appearance relighting. We further compare the trained 2D prior against other previous approaches that predict albedo and material from a single image using test data from BlenderVault excluded from training.

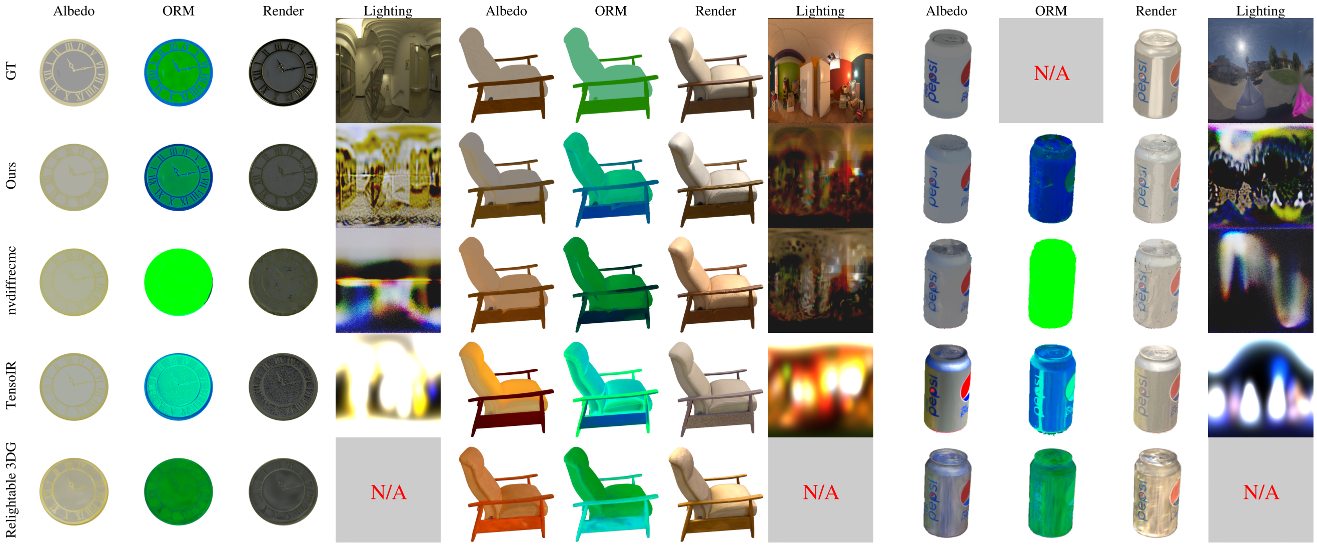


Figure 4. Qualitative comparison for MaterialFusion vs. other methods. We present the 3D reconstructed albedo, ORM, environment light map, and relit rendered images for three different objects, both synthetic and real. Our method demonstrates better accuracy compared to the baseline methods, as can be seen by the accuracy of the reconstructed materials and the relit image appearance. Our prior also acts as an additional regularizer on other 3D properties such as geometry and illumination.

4.1. Relightable 3D Reconstruction

For MaterialFusion, we adopt a validation setup similar to nvdiffrecmc by relighting the objects under novel illuminations and then comparing to the groundtruth relit object images. For synthetic objects, we acquire the groundtruth relightings by rendering images of synthetic objects under unseen illuminations, while real objects are captured in novel environments for which the illuminations are computed.

Datasets. We use 4 objects from NeRFactor [58], 5 objects from the NeRF synthetic dataset [32], 9 test objects from BlenderVault, and 14 objects from the Stanford-ORB [23] datasets. The first three datasets consist of diverse synthetic objects with camera poses and their groundtruth data allows for us to re-render and compare objects with different illuminations. The NeRFactor and BlenderVault objects are relit by eight low resolution environment maps while NeRF synthetic objects are relit by four high resolution environment maps, and the quality comparison is computed on a test set of eight unseen poses per environment map. We also show our relighting performance on real objects from the Stanford-ORB dataset, which has images, corresponding poses, and groundtruth illuminations allowing us to re-render and relight objects. Objects are relit under two novel illuminations and the relighting comparison is done using a test set of unseen poses per environment map.

Metrics. The final results for the 3D pipeline relighting comparison are the average PSNR, SSIM, and LPIPS across all relighting test views for each dataset. The metrics for the albedo used were PSNR, SSIM, and L1, and PSNR and L1 for ORM. LPIPS was excluded for both since perceptual similarity does not matter for ORM, and the VGG network likely has not seen albedo images. Given the scaling ambiguity between the albedo and light intensity during inference, the channels of RGB and albedo images are scaled against groundtruth during validation [17]. Since the ORM and albedo are fundamentally pixel-wise material parameters, we use the L1 metric to measure physical similarity.

Baselines. We compare MaterialFusion against three current state-of-the-art inverse rendering methods that estimate geometry, albedo, roughness, and metallicness from a set of images. These approaches are nvdiffrecmc [17], which is the method our pipeline is built upon, Relightable 3D Gaussian [14], and TensoIR [20].

Results. We present qualitative and quantitative results for both albedo and ORM estimation quality as well as performance during relighting. Fig. 4 shows a visual comparison of the albedo and ORM estimated by all methods. Our method is able to recover high frequency details in both the albedo and ORM that other methods are not able to. This in turn leads to better performance under novel relighting, where Tab. 1 shows our method achieving the highest scores across all three datasets. The source of improvement in the relighting performance is best understood via the results in Tab. 2, where the estimated albedo and ORM quality for BlenderVault objects were directly compared to the groundtruth. We were unable to compare for the NeRF and

	NeRF Synthetic			NeRFactor			BlenderVault			Stanford-ORB		
	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
nvdiffrecmc	25.70	0.924	0.090	25.91	0.921	0.092	25.11	0.910	0.167	31.10	0.968	0.048
TensoIR	24.32	0.923	0.090	24.87	0.916	0.094	25.01	0.903	0.162	28.81	0.959	0.047
Relightable3DG	23.08	0.897	0.094	23.99	0.908	0.082	22.83	0.909	0.148	27.40	0.955	0.048
MaterialFusion	26.26	0.927	0.085	26.31	0.922	0.091	26.33	0.921	0.143	31.68	0.967	0.046

Table 1. Comparison of novel view synthesis relighting. In each column, the best , second best , and third best results are marked.

	Albedo			ORM	
	PSNR ↑	SSIM ↑	L1 ↓	PSNR ↑	L1 ↓
nvdiffrecmc	28.24	0.939	0.021	15.93	0.062
TensoIR	25.82	0.927	0.026	14.19	0.063
Relightable3DG	24.30	0.925	0.036	20.96	0.041
MaterialFusion	29.31	0.949	0.015	22.21	0.033

Table 2. Reconstructed 3D albedo and ORM comparison on BlenderVault objects.

NeRFactor datasets as the underlying material shaders used did not conform directly to the albedo and ORM.

The comparisons show that our method performs best in both albedo and ORM estimation, as other methods suffer from poor albedo or ORM estimates, leading to poorer relighting. The synthetic example of the clock in Fig. 4 shows how MaterialFusion is able to accurately disambiguate different areas of the material and albedo, leading to a much more accurate rendering under novel illumination where details aren't lost like in the other methods' renderings. This can also be seen in the real can example where our method accurately deduces it is metallic (shown by the strength of the blue channel in the ORM map) and is able to accurately replicate the reflection of the can similarly to the real world. Our method shows better semantic material understanding as it is able to correctly distinguish between different parts of an object that are made of different materials, leading to better decoupling between the reflectance and environment illumination. Our results confirm our prior's improvements against baselines by better inferring underlying physical properties on synthetic and real data.

4.2. Validating Material Inference from a 2D Input

To validate StableMaterial's performance, we evaluate its performance on RGB images of synthetic test objects captured under an unknown illumination. We then directly compare the predictions to the groundtruth albedo and materials using extracted data from the synthetic objects.

Datasets. We utilize 8 diverse test objects from our BlenderVault datasets whose material data was not seen during training. We then render the groundtruth albedo, ORM, and RGB appearance under an unknown randomly selected fixed illumination for 4 views. This is done per each object.

	Albedo			ORM	
	PSNR ↑	SSIM ↑	L1 ↓	PSNR ↑	L1 ↓
Derender3D	21.69	0.874	0.025	–	–
IIR	22.62	0.905	0.026	18.68	0.041
IID	21.71	0.871	0.031	18.87	0.045
StableMaterial	24.25	0.902	0.018	24.86	0.029
StableMaterialMV	24.70	0.907	0.018	26.34	0.014

Table 3. Comparison of albedo and ORM 2D predictions produced by our method versus other methods. We use the mean of 10 samples generated by StableMaterial and StableMaterialMV for the evaluation, where StableMaterialMV denotes our prior utilizing multi-view attention during inference.

Metrics. To account for the variance in StableMaterial's outputs, we follow the procedure described in [22] and compute 10 estimates for the albedo and ORM images and average them together before comparing to the groundtruth. We further account for the scale ambiguity in the resulting albedo for all the baselines by rescaling to the groundtruth albedo. Similarly to the 3D evaluation, the metrics used for the albedo were PSNR, SSIM, and L1, and PSNR and L1 for ORM. The final results are computed as the mean across views for all objects.

Baselines. We test StableMaterial against Inverse Indoor Rendering (IIR) [62] and Intrinsic Image Diffusion (IID) [22], which were trained on scene data to directly predict the albedo, roughness, metallicness given a single image. We also include [52], which was trained on diverse data and predicts the albedo but not materials.

Multi-view Attention at Inference. To make StableMaterial produce material outputs consistent across 2D views, we follow previous works [45, 50] and incorporate multi-view attention. Specifically, we input a batch of 4 images, and modify the self-attention layers of the model so that each latent pixel in each of the images attends to the latent pixels of all other images. As such, StableMaterial predicts the most likely material given all input image appearances. The self-attention layers of the network process the 4 images as a single large image, while the other layers process them independently. Importantly, this multi-view attention mechanism is only employed during inference for 2D images. We show that using multi-view attention improves the quality of inferred materials against other baselines.

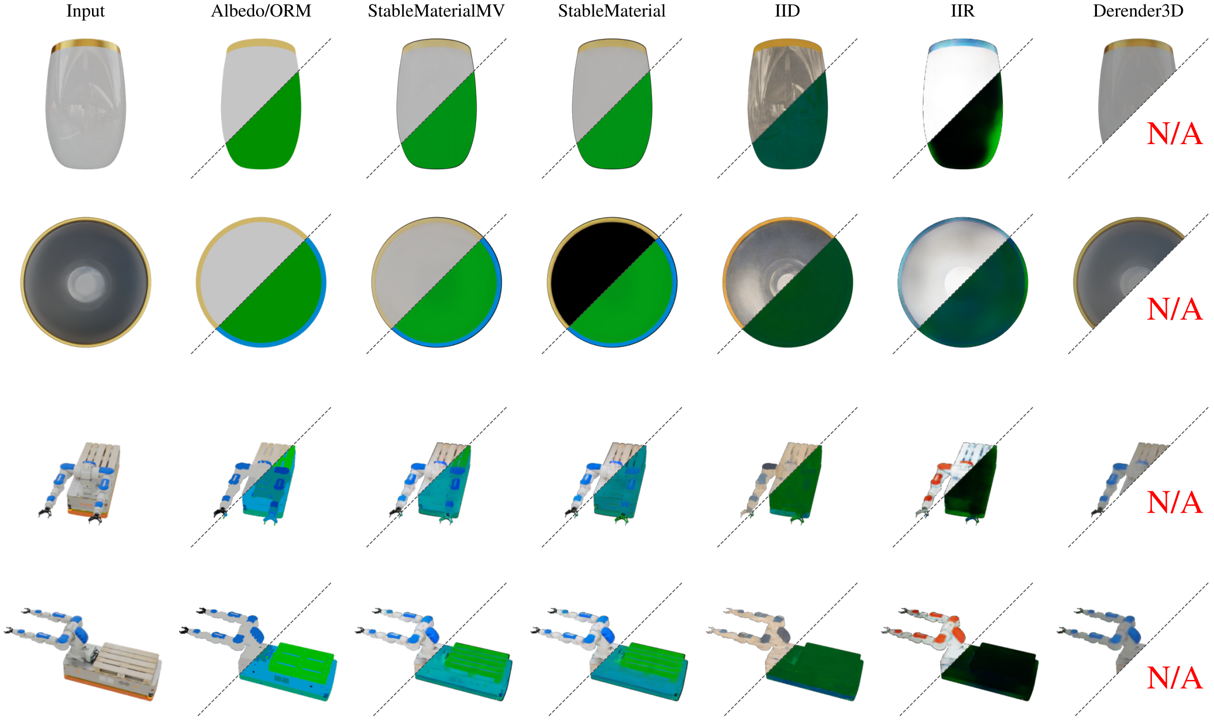


Figure 5. Qualitative comparison of the albedo and ORM 2D predictions. The Derender3D ORM data is marked as N/A since it does not offer ORM predictions. Given 4 images of an object, StableMaterial recovers complex material data. StableMaterialMV attends to appearance details across views, recovering consistent and high quality materials across challenging views, as seen in the cup example.

	BlenderVault		
	PSNR ↑	SSIM ↑	LPIPS ↓
$\gamma_i = 1$	25.50	0.916	0.150
$\lambda_{\text{RGB}} = 0$	21.41	0.871	0.233
$\lambda_{\text{latent}} = 0$	26.12	0.917	0.147
Ours	26.33	0.921	0.143

Table 4. Effects of ablating elements from $\mathcal{L}_{\text{MaterialFusion}}$.

Results. As shown in Tab. 3, our trained model shows strong performance across multiple objects in estimating the Albedo and ORM quality. Notably, performance jumps further when multi-attention is used across 4 views, as our model can handle difficult views that offer little appearance information during inference. Fig. 5 shows a qualitative comparison of our model against previous approaches.

The consistent improvements in both 2D and 3D tasks highlight the effectiveness of our approach in capturing the underlying physical properties of objects. The multi-view variant further demonstrates the benefits of leveraging additional viewpoints to enhance the albedo and ORM prediction quality for difficult 2D views. We found no significant differences when employing multi-view attention during inverse rendering, given that the albedo and ORM representations are already optimized to be multi-view consistent. However, the performance boost for 2D images raises the potential for usage in sparse view scenarios or where 3D reconstruction is not needed or infeasible.

4.3. Ablation Studies

In Tab. 4, we ablate three terms of $\mathcal{L}_{\text{MaterialFusion}}$ and evaluate relighting performance on the BlenderVault test dataset. All other parameters are unchanged when ablating one parameter. Setting $\lambda_{\text{RGB}} = 0$ particularly affects performance; by backpropagating through the SD encoder, the latent SDS term gradient introduces artifacts in the materials, degrading their quality. We also conduct an ablation where γ_i is set to 1 throughout the inverse rendering optimization. This leads to a noticeable drop in performance, as materials estimated for objects with finer details suffer.

5. Conclusion

In this paper, we introduced MaterialFusion, a 3D inverse rendering approach that utilizes StableMaterial, a 2D diffusion model finetuned from Stable Diffusion as a prior for enhancing the underlying materials during training. We introduced BlenderVault, a dataset of high quality objects and underlying PBR assets used to finetune our prior, enabling it with knowledge to recreate complex materials from images. Utilizing our prior on top of an off the shelf inverse rendering approach lead to a significant performance boost when for inferring relightable 3D representations. While our work introduces distillation of material knowledge in a 3D scenario, we believe there is great potential in utilizing our prior for applications in 2D or sparse-view settings.

6. Acknowledgments

The authors would like to thank Jianjin Xu and the rest of the students from the Human Sensing Laboratory for their helpful feedback. This work was supported in part by the NSF GFRP (Grant No. DGE1745016) and NSF Award IIS-2345610.

References

[1] 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

[2] 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

[3] Sai Bi, Zexiang Xu, Kalyan Sunkavalli, David Kriegman, and Ravi Ramamoorthi. Deep 3d capture: Geometry and reflectance from sparse multi-view images. In *CVPR*, 2020. 2

[4] 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

[5] Aljaž Božič, Denis Gladkov, Luke Doukakis, and Christoph Lassner. Neural assets: Volumetric object capture and rendering for interactive environments. *arxiv*, 2022. 2

[6] Tim Brooks, Aleksander Holynski, and Alexei A. Efros. Instructpix2pix: Learning to follow image editing instructions. In *CVPR*, 2023. 4

[7] Brent Burley and Walt Disney Animation Studios. Physically-based shading at disney. In *SIGGRAPH*. ACM Transactions on Graphics (ToG), 2012. 3

[8] Rui Chen, Yongwei Chen, Ningxin Jiao, and Kui Jia. Fantasia3d: Disentangling geometry and appearance for high-quality text-to-3d content creation. In *ICCV*, 2023. 2

[9] Zhaoxi Chen, Gyeongsik Moon, Kaiwen Guo, Chen Cao, Stanislav Pidhorskyi, Tomas Simon, Rohan Joshi, Yuan Dong, Yichen Xu, Bernardo Pires, He Wen, Lucas Evans, Bo Peng, Julia Buffalini, Autumn Trimble, Kevyn McPhail, Melissa Schoeller, Shoou-I Yu, Javier Romero, Michael Zollhöfer, Yaser Sheikh, Ziwei Liu, and Shunsuke Saito. Urhand: Universal relightable hands. In *CVPR*, 2024. 2

[10] Ziang Cheng, Junxuan Li, and Hongdong Li. Wildlight: In-the-wild inverse rendering with a flashlight. *CVPR*, 2023. 2

[11] Jasmine Collins, Shubham Goel, Kenan Deng, Achleshwar Luthra, Leon Xu, Erhan Gundogdu, Xi Zhang, Tomas F Yago Vicente, Thomas Dideriksen, Himanshu Arora, et al. Abo: Dataset and benchmarks for real-world 3d object understanding. In *CVPR*, 2022. 3

[12] 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. 3

[13] Laura Downs, Anthony Francis, Nate Koenig, Brandon Kinman, Ryan Hickman, Krista Reymann, Thomas B. McHugh, and Vincent Vanhoucke. Google scanned objects: A high-quality dataset of 3d scanned household items. *arxiv*, 2022. 3

[14] 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, 6

[15] 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. 4

[16] Yuan-Chen Guo, Di Kang, Linchao Bao, Yu He, and Song-Hai Zhang. Nerfren: Neural radiance fields with reflections. In *CVPR*, 2022. 2

[17] Jon Hasselgren, Nikolai Hofmann, and Jacob Munkberg. Shape, Light, and Material Decomposition from Images using Monte Carlo Rendering and Denoising. *NeurIPS*, 2022. 2, 3, 5, 6

[18] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NeurIPS*, 2020. 2

[19] 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. In *NeurIPS*, 2023. 3

[20] 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, 6

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

[22] Peter Kocsis, Vincent Sitzmann, and Matthias Nießner. Intrinsic image diffusion for indoor single-view material estimation. *CVPR*, 2024. 3, 7

[23] Zhengfei Kuang, Yunzhi Zhang, Hong-Xing Yu, Samir Agarwala, Elliott Wu, Jiajun Wu, et al. Stanford-orb: A real-world 3d object inverse rendering benchmark. *NeurIPS*, 2023. 2, 6

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

[25] Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler, Ming-Yu Liu, and Tsung-Yi Lin. Magic3d: High-resolution text-to-3d content creation. In *CVPR*, 2023. 2

[26] 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. *CVPR*, 2024. 3

[27] Minghua Liu, Chao Xu, Haian Jin, Linghao Chen, Mukund Varma T, Zexiang Xu, and Hao Su. One-2-3-45: Any single image to 3d mesh in 45 seconds without per-shape optimization. *NeurIPS*, 2024.

[28] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object. *ICCV*, 2023. 2

[29] Yuan Liu, Cheng Lin, Zijiao Zeng, Xiaoxiao Long, Lingjie Liu, Taku Komura, and Wenping Wang. Syncdreamer: Generating multiview-consistent images from a single-view image. *ICLR*, 2024. 2

[30] 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. *CVPR*, 2024. 3

[31] Gal Metzer, Elad Richardson, Or Patashnik, Raja Giryes, and Daniel Cohen-Or. Latent-nerf for shape-guided generation of 3d shapes and textures. In *CVPR*, 2023. 2

[32] 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. 2, 6

[33] Piotr Mirowski, Andras Banki-Horvath, Keith Anderson, Denis Teplyashin, Karl Moritz Hermann, Mateusz Malinowski, Matthew Koichi Grimes, Karen Simonyan, Koray Kavukcuoglu, Andrew Zisserman, et al. The streetlearn environment and dataset. *NeurIPS*, 2018. 4

[34] 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

[35] 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, 3, 5

[36] Keunhong Park, Konstantinos Rematas, Ali Farhadi, and Steven M. Seitz. Photoshape: photorealistic materials for large-scale shape collections. *ACM Transactions on Graphics (ToG)*, 2018. 3

[37] Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *ICLR*, 2023. 2, 3, 5

[38] Jeremy Reizenstein, Roman Shapovalov, Philipp Henzler, Luca Sbordone, Patrick Labatut, and David Novotny. Common objects in 3d: Large-scale learning and evaluation of real-life 3d category reconstruction. In *ICCV*, 2021. 3

[39] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. *CVPR*, 2022. 2, 3, 4

[40] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. *NeurIPS*, 2022. 2

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

[42] Sam Sartor and Pieter Peers. Matfusion: A generative diffusion model for svbrdf capture. In *SIGGRAPH Asia*. ACM, 2023. 3

[43] Tianchang Shen, Jun Gao, Kangxue Yin, Ming-Yu Liu, and Sanja Fidler. Deep marching tetrahedra: a hybrid representation for high-resolution 3d shape synthesis. In *NeurIPS*, 2021. 5

[44] Ruoxi Shi, Hansheng Chen, Zhuoyang Zhang, Minghua Liu, Chao Xu, Xinyue Wei, Linghao Chen, Chong Zeng, and Hao Su. Zero123++: a single image to consistent multi-view diffusion base model. *arxiv*, 2023. 3

[45] Yichun Shi, Peng Wang, Jianglong Ye, Mai Long, Kejie Li, and Xiao Yang. Mvdream: Multi-view diffusion for 3d generation. *ICLR*, 2024. 2, 3, 7

[46] Jingxiang Sun, Bo Zhang, Ruizhi Shao, Lizhen Wang, Wen Liu, Zhenda Xie, and Yebin Liu. Dreamcraft3d: Hierarchical 3d generation with bootstrapped diffusion prior. In *ICLR*, 2024. 2

[47] Jiaxiang Tang, Zhaoxi Chen, Xiaokang Chen, Tengfei Wang, Gang Zeng, and Ziwei Liu. Lgm: Large multi-view gaussian model for high-resolution 3d content creation. *ECCV*, 2024. 3

[48] Giuseppe Vecchio, Rosalie Martin, Arthur Roullier, Adrien Kaiser, Romain Rouffet, Valentin Deschaintre, and Tamy Boubekeur. Controlmat: A controlled generative approach to material capture. *arxiv*, 2023. 3

[49] Haochen Wang, Xiaodan Du, Jiahao Li, Raymond A Yeh, and Greg Shakhnarovich. Score jacobian chaining: Lifting pretrained 2d diffusion models for 3d generation. In *ICCV*, 2023. 2

[50] Peng Wang and Yichun Shi. Imagedream: Image-prompt multi-view diffusion for 3d generation. *ICLR*, 2024. 2, 3, 7

[51] Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. In *NeurIPS*, 2023. 2

[52] Felix Wimbauer, Shangzhe Wu, and Christian Rupprecht. De-rendering 3d objects in the wild. *CVPR*, 2022. 7

[53] Guanjun Wu, Taoran Yi, Jiemin Fang, Lingxi Xie, Xiaopeng Zhang, Wei Wei, Wenyu Liu, Qi Tian, and Xinggang Wang. 4d gaussian splatting for real-time dynamic scene rendering. In *CVPR*, 2024. 2

[54] 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. *CVPR*, 2024. 3

[55] 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. *arxiv*, 2024. 3

[56] Zheng Zeng, Valentin Deschaintre, Iliyan Georgiev, Yannick Hold-Geoffroy, Yiwei Hu, Fujun Luan, Ling-Qi Yan, and Miloš Hašan. Rgb↔x: Image decomposition and synthesis using material- and lighting-aware diffusion models. In *SIGGRAPH*. ACM, 2024. 3

[57] Kai Zhang, Fujun Luan, Zhengqi Li, and Noah Snavely. Iron: Inverse rendering by optimizing neural sdfs and materials from photometric images. In *CVPR*, 2022. 2

[58] 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, 6

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

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

[61] Junzhe Zhu and Peiye Zhuang. Hifa: High-fidelity text-to-3d generation with advanced diffusion guidance. In *ICLR*, 2024. 5

[62] Jingsen Zhu, Fujun Luan, Yuchi Huo, Zihao Lin, Zhihua Zhong, Dianbing Xi, Rui Wang, Hujun Bao, Jiaxiang Zheng, and Rui Tang. Learning-based inverse rendering of complex indoor scenes with differentiable monte carlo raytracing. In *SIGGRAPH Asia*. ACM, 2022. 7