

# Neural LightRig: Unlocking Accurate Object Normal and Material Estimation with Multi-Light Diffusion

Zexin He<sup>1\*</sup>, Tengfei Wang<sup>2\*</sup>, Xin Huang<sup>2</sup>, Xingang Pan<sup>3</sup>, Ziwei Liu<sup>3</sup>

<sup>1</sup>The Chinese University of Hong Kong, <sup>2</sup>Shanghai AI Lab, <sup>3</sup>Nanyang Technological University

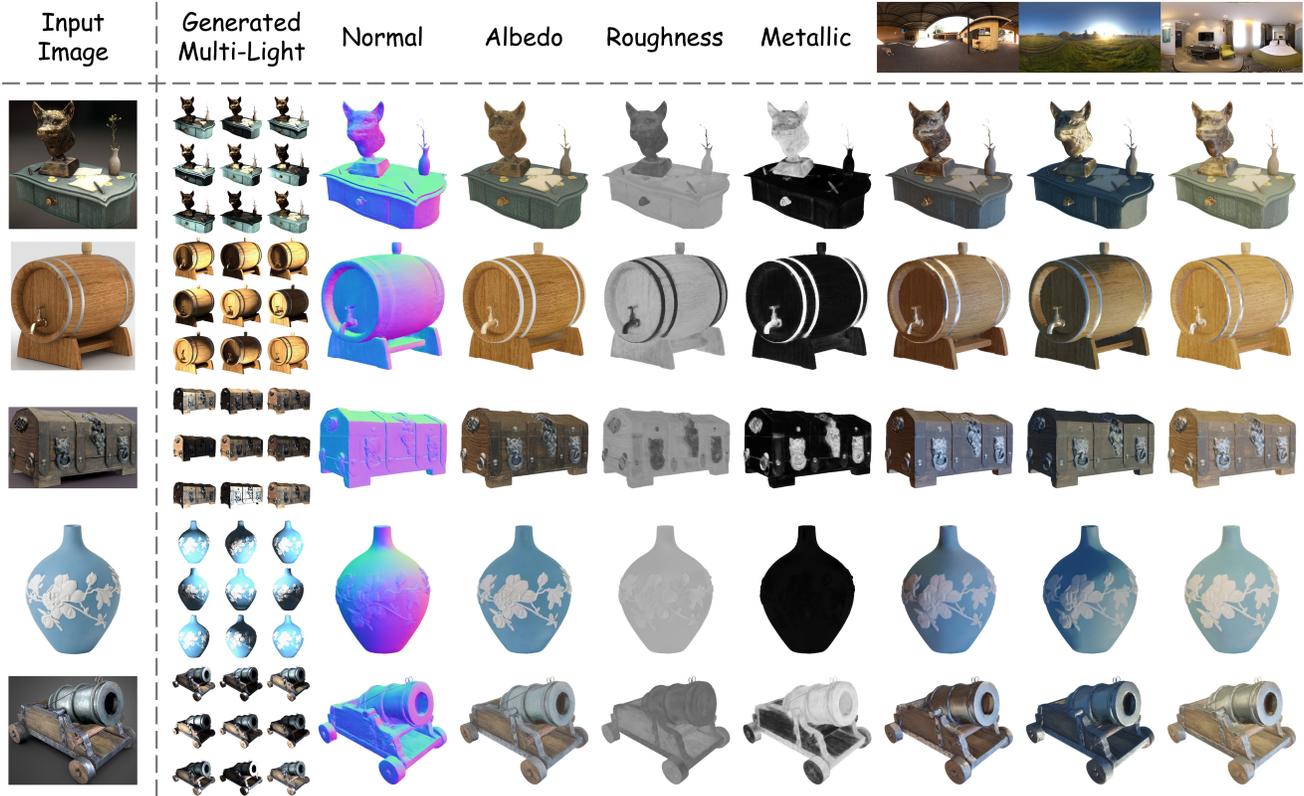


Figure 1. *Neural LightRig* takes an image as input and generates multi-light images to assist the estimation of high-quality normal and PBR materials, which can be used to render realistic relit images under various environment lighting.

## Abstract

Recovering the geometry and materials of objects from a single image is challenging due to its under-constrained nature. In this paper, we present **Neural LightRig**, a novel framework that boosts intrinsic estimation by leveraging auxiliary multi-lighting conditions from 2D diffusion priors. Specifically, **1)** we first leverage illumination priors from large-scale diffusion models to build our multi-light diffusion model on a synthetic relighting dataset with dedicated designs. This diffusion model generates multiple consistent images, each illuminated by point light sources

in different directions. **2)** By using these varied lighting images to reduce estimation uncertainty, we train a large G-buffer model with a U-Net backbone to accurately predict surface normals and materials. Extensive experiments validate that our approach significantly outperforms state-of-the-art methods, enabling accurate surface normal and PBR material estimation with vivid relighting effects. Code and dataset are available on our project page at <https://projects.zxhezexin.com/neural-lightrig>.

## 1. Introduction

Recovering the geometry and physically-based rendering (PBR) materials of real-world objects from images is a piv-

\* Equal contribution. Work done during Zexin He’s internship at Shanghai AI Lab.

total problem in graphics and computer vision. This task, also known as inverse rendering, facilitates a wide range of applications, such as video gaming, augmented and virtual reality, and robotics. In this paper, we proposed a data-driven approach for jointly estimating the surface normal and PBR materials of objects from a single image. Due to the complex interaction among geometry, materials, and environmental lighting, this ill-posed problem remains particularly challenging.

Prior research [6, 17] has predominantly focused on optimization-based generation through differentiable rendering, which compares forward-rendered images with input images to refine normals and PBR materials. However, these methods are often time-consuming and heavily reliant on the capabilities of the differentiable renderer [27]. Though some works explored feed-forward estimation [34, 54, 57], their quality and generalizability still remain challenging, due to the inherently ill-posed nature of inferring geometry and materials from a single image.

For precise normal and material acquisition, photometric stereo techniques [51] are widely employed, as they mitigate ambiguity by capturing multiple images from the same viewpoint with various lighting. These images are illuminated by different point light sources, which provide variations in surface reflectance to enrich information. However, such methods [10, 13, 28] often require complex capture systems with sophisticated cameras or lighting setups, which can be costly and impractical for in-the-wild images. Given the promising advances in image diffusion models, we ask the question: can we develop a multi-light diffusion model to simulate images illuminated by different directional light sources, thereby improving surface normal and material estimation (as shown in Fig. 1)?

Our motivation arises from recent advances in 3D generation, which employ diffusion models [30, 43] to generate multi-view images and train reconstruction models [21] for 3D reconstruction. These multi-view diffusion models have demonstrated the potential to manipulate camera views of pre-trained image diffusion models such as Stable Diffusion [38]. Similarly, we aim to expand the use of pre-trained diffusion models for multi-light image generation.

In this work, we present *Neural LightRig* for joint normal and material estimation of objects from monocular images, which consists of a multi-light diffusion model and a large prediction model. Given an input image, the **multi-light diffusion model** produces consistent and high-quality relit images under various point light sources (as shown in Fig. 4). To achieve this, we create a synthetic relighting dataset for training with Blender [9]. With a dedicated architecture and training design, our diffusion model enables the multi-light generation of objects from arbitrary categories. The **large G-buffer model** then processes the generated multi-light images to produce surface normals and

PBR materials, such as albedo, roughness, and metallic. We employ a UNet architecture for efficient and high-resolution prediction, with end-to-end supervision at the pixel level. To bridge the domain gap between multi-light images rendered from 3D objects and those generated by diffusion models, we further design a series of data augmentation strategies for domain alignment.

Taken together, the proposed framework demonstrates remarkable performance on both synthetic and real-world images. Extensive qualitative and quantitative evaluations show that *Neural LightRig* surpasses existing approaches in surface normal estimation, PBR material estimation, and single-image relighting. Comprehensive visual results are provided in the appendix and on our [project page](#). Our key contributions are as follows:

- We propose a novel approach for object normal and PBR estimation from monocular images, reformulating this ill-posed problem by simulating multi-lighting conditions.
- We construct a synthetic dataset for multi-light image generation and surface property estimation. With this dataset, we demonstrate the capability to manipulate diffusion models for consistent multi-light generation.
- Extensive experiments validate the effectiveness of our method, establishing new state-of-the-art results.

## 2. Related Works

**Diffusion Models.** Well-trained diffusion models [38, 49] have shown promising potential in providing essential priors for under-determined tasks. Recent works showcase the utility of image diffusion models in novel-view synthesis [32, 33, 35, 43, 44, 50], which combines with reconstruction models [18, 21, 46] to achieve high-quality 3D generation. Similarly, some recent works attempt to leverage the learned priors in diffusion models to simulate lighting variations [23, 56], but they do not account for the consistency of multi-light generation. In contrast, we aim to generate multiple images under different lighting sources that facilitate object surface property estimation.

**Monocular Normal Estimation.** Estimating surface normals from a single image is a classic yet under-determined problem. Early works often relied on photometric cues or handcrafted features [15, 19, 20], while later works adopted deep learning to improve accuracy [4, 12, 26, 29, 37, 48, 52, 62]. More recently, large-scale datasets [11, 14] have further advanced regression-based methods [2, 3, 5]. Despite promising results, they struggle with complex details due to inherent ambiguity. Diffusion-based methods [16, 25, 53], turn to generative priors [38] to help address such ambiguity but often fall short in accurately aligning with ground truth, leading to deviations in finer geometric details crucial for downstream tasks.

**Material Estimation.** Material estimation aims to recover intrinsic properties from images, which is an ill-posed prob-



Figure 2. **Framework Overview.** Multi-light diffusion generates multi-light images from an input image. These images with corresponding lighting orientations are then used to predict surface normals and PBR materials with a regression U-Net.

lem, as multiple combinations of materials and lighting conditions could lead to the same appearance, Traditional methods attempted to employ photometric stereotypes [13, 51] to disambiguate this problem under controlled lighting conditions [10, 28]. Some works [7, 17, 45, 58] optimize neural representation with multi-view images. Later, the emergence of large-scale synthetic datasets [11, 47] has advanced data-driven approaches [31, 34, 41, 42, 54, 55], but they still contend with under-determination. Recently, diffusion-based methods [8, 22, 36, 57] have emerged as a promising alternative, but often suffer from domain shift between material images and natural images.

### 3. Approach

Given an image  $\mathbf{I}$ , we aim to estimate both its surface normal  $\mathbf{n}$  and PBR materials (albedo  $\mathbf{a}$ , roughness  $\mathbf{r}$ , and metallic  $\mathbf{m}$ ), where  $\mathbf{n}, \mathbf{a} \in \mathbb{R}^{H \times W \times 3}$  and  $\mathbf{r}, \mathbf{m} \in \mathbb{R}^{H \times W \times 1}$ . These surface properties, commonly known as G-buffers in graphics, are collectively denoted as  $\mathcal{B} = \{\mathbf{n}, \mathbf{a}, \mathbf{r}, \mathbf{m}\}$ . However, interpreting these properties from a single lighting condition is challenging due to the under-constrained nature of the problem. To address this, we propose *Neural LightRig*, as illustrated in Fig. 2. Our approach leverages a multi-light diffusion model (Sec. 3.1) to generate multi-light images from the input, which then act as enriched conditions to alleviate the inherent ambiguity in G-buffer prediction model (Sec. 3.2). We further describe the construction of our synthetic dataset, *LightProp*, which supports both stages of our framework, in Sec. 3.3.

#### 3.1. Multi-Light Diffusion

To obtain surface reflectance variations that increase contextual information for accurate G-buffer estimation, we learn a diffusion model  $g(\cdot)$  to generate  $L$  multi-light images from the input image  $\mathbf{I}$ :

$$\{\mathbf{x}^i \mid i = 1, 2, \dots, L\} = g(\mathbf{I}). \quad (1)$$

In particular, we set  $L = 9$  to balance performance and efficiency, covering a diverse range of lighting variations

(Fig. 4) without excessive overhead.

**Generating Multi-Light Images.** Collecting such training pairs is challenging due to the limited availability of 3D objects with PBR [11, 47] and the high cost of real-world capturing in photometric stereotypes [24]. Fortunately, diffusion models trained on massive internet images have shown an inherent ability to model complex 3D shapes and textures, which have been applied for novel view synthesis [43] and relighting [23, 56]. We thus leverage the prior from a well-trained image diffusion model and fine-tune it for multi-light generation, arguing that such a well-trained image generation model possesses the capacity to simulate diverse lighting conditions. Rather than generating each-light image  $\mathbf{x}^i$  separately, we arrange nine-light images in a  $3 \times 3$  grid layout to form a single image  $\mathbf{x}$ , allowing the simultaneous generation for them. This simple configuration facilitates efficient cross-image context communication, thereby enhancing the consistency of generated multi-light images.

**Conditioning Strategy.** To incorporate the input image into the diffusion model, we employ a hybrid conditioning method, as illustrated in Fig. 3. As the input images are pixel-wise aligned with the multi-light images, we naturally apply *channel-wise concatenation*. This straightforward concatenation effectively captures the variations between the input and each multi-light image, which is essential for generating accurate lighting effects. However, we found this simple concatenation alone is inadequate for generating high-fidelity multi-light images, leading to discrepancies in color tone and texture relative to the input. To address this, we further adopt *reference attention* [43, 59], where self-attention layers in the denoising U-Net also attend to keys and values obtained from the input image. This is represented as  $\text{Attn}(\mathbf{Q}, [\mathbf{K}, \mathbf{K}_{\text{cond}}], [\mathbf{V}, \mathbf{V}_{\text{cond}}])$ , in which  $\mathbf{Q}, \mathbf{K}, \mathbf{V}$  are the query, key, and value tokens from the denoising stream, and the subscript “cond” denotes tokens from the input image. This combined approach manages to preserve desired textures from in the input and is crucial for generating high-quality and realistic multi-light images.

**Tuning Scheme.** We build our model on Stable Diffusion

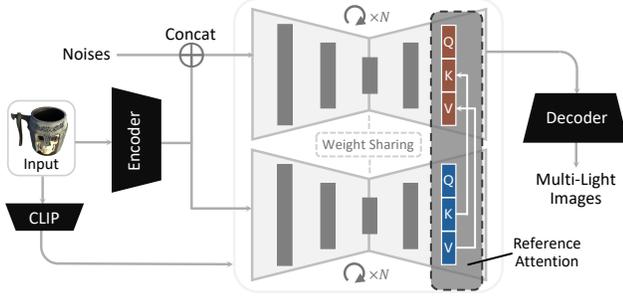


Figure 3. Hybrid condition in multi-light diffusion. Input images are incorporated via *concatenation* with noise latents and enhanced through *reference attention*, where queries in the denoise stream attend to keys and values from both streams.

$v$ -version model [1, 40]. Let  $\alpha_t, \sigma_t$  be the controlling factors in the diffusion process, and define ground-truth velocity as  $\mathbf{v} = \alpha_t \epsilon + \sigma_t \mathbf{x}$  and predicted velocity as  $\mathbf{v}_\theta(\cdot)$ . The training target can be denoted as:

$$\mathcal{L} = \mathbb{E}_{\mathbf{x}, \mathbf{I}, \epsilon, t} [\|\mathbf{v} - \mathbf{v}_\theta(\mathbf{z}_t, t, \mathbf{I})\|^2], \quad (2)$$

where  $\mathbf{z}_t$  is the noisy latent of  $\mathbf{x}$  at timestep  $t$ , and  $\mathbf{I}$  is the input image. To fully leverage the capacity of diffusion model, we adopt a two-phase training scheme. Initially, we freeze most parameters except for the first convolution layer and all attention layers to warm up the weights. This stabilizes early training, allowing for a smooth transition without severely disrupting the pre-trained model. Afterwards, we fine-tune the entire model at a considerably lower learning rate, facilitating careful adaptation for multi-light generation while retaining as much prior knowledge as possible.

### 3.2. Large G-Buffer Model

Next, we learn a regression model  $f(\cdot)$  to predict normals and PBR maps with the auxiliary multi-light images.

**Prediction Model.** Since the input image, multi-light images, and G-buffer maps are pixel-wise aligned, we opt for a U-Net architecture thanks to its efficiency in high-resolution prediction. Also, U-Net provides inductive bias for learning spatial relations, making it well-suited for our task. The model takes channel-wise concatenated input and multi-light images, and outputs an 8-channel G-buffer, containing 3-channel  $\mathbf{n}$  and  $\mathbf{a}$  maps, and 1-channel  $\mathbf{r}$  and  $\mathbf{m}$  maps. This multi-light-enhanced G-buffer prediction is represented as:

$$\mathbf{B} = f(\mathbf{I}, \{(\mathbf{x}^i, \theta^i, \varphi^i) \mid i = 1, 2, \dots, L\}), \quad (3)$$

where each novel-light image  $\mathbf{x}^i$  is associated with the light source poses  $\theta^i$  and  $\varphi^i$ , which indicate spherical coordinates of the light source relative to the object (see Fig. 4). Conditioning on these poses allows  $f(\cdot)$  to explicitly correlate shading variations with their respective light sources, enhancing surface estimation.

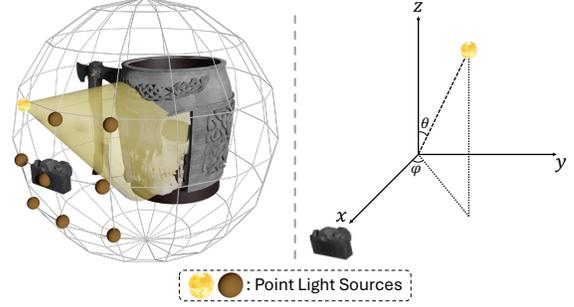


Figure 4. Visualization of multi-light setup in *LightProp*. Camera and point lights are positioned on a sphere around the object.  $\theta, \varphi$  are spherical coordinates to determine each light’s orientation relative to the object.

**Training Objectives.** To train the model  $f(\cdot)$  for G-buffer prediction, we apply loss functions to each of the G-buffer properties. We employ a cosine similarity loss for normals, enforcing the model to capture precise surface orientations. To stabilize the training, we also include an MSE term as regularization:

$$\mathcal{L}_{\text{normal}} = \left(1 - \frac{\mathbf{n} \cdot \hat{\mathbf{n}}}{\|\mathbf{n}\| \|\hat{\mathbf{n}}\|}\right) + \lambda_1 \|\mathbf{n} - \hat{\mathbf{n}}\|^2, \quad (4)$$

where  $\hat{\mathbf{n}}$  and  $\mathbf{n}$  are the predicted and ground-truth normals. For the predicted albedo  $\hat{\mathbf{a}}$ , roughness  $\hat{\mathbf{r}}$ , and metallic  $\hat{\mathbf{m}}$ , we simply use MSE losses as:

$$\mathcal{L}_{\text{PBR}} = \|\mathbf{a} - \hat{\mathbf{a}}\|^2 + \|\mathbf{r} - \hat{\mathbf{r}}\|^2 + \|\mathbf{m} - \hat{\mathbf{m}}\|^2. \quad (5)$$

The overall loss is the weighted sum of the two losses.

**Augmentations.** We train our prediction model using ground-truth rendered multi-light images, but for inference, we rely on generated images from diffusion models. In our earlier experiments, we observed a domain gap between the generated and rendered multi-light images in sharpness and brightness. This gap would introduce discrepancies between training and inference, causing degraded performances. To bridge this gap, we apply a series of augmentations to multi-light images during training, including: (a) *Random Degradation*, such as resizing and grid distortion that simulate small misalignments; (b) *Random Intensity* that adjusts brightness in HSV space, simulating brightness variations of multi-light images; (c) *Random Orientation* perturbs  $\{\theta^i, \varphi^i\}$  to account for potential disparities, encouraging  $f(\cdot)$  to be robust to inaccurate lighting cues; and (d) *Data Mixing*, where we mix generated multi-light images into the training data to further mitigate this gap.

### 3.3. LightProp Dataset

To train our model, we need to collect paired multi-light images and corresponding normal and PBR material maps.

Table 1. Quantitative comparison on surface normal estimation. We report mean and median angular errors, as well as accuracies within different angular thresholds from 3° to 30°.

Method	Mean ↓	Median ↓	3° ↑	5° ↑	7.5° ↑	11.25° ↑	22.5° ↑	30° ↑
RGB↔X [57]	14.847	13.704	11.676	23.073	35.196	49.829	75.777	86.348
DSINE [2]	9.161	7.457	23.565	41.751	57.596	72.003	90.294	95.297
GeoWizard [16]	8.455	6.926	22.245	40.993	58.457	74.916	93.315	97.162
Marigold [25]	8.652	7.078	25.219	42.289	58.062	72.873	92.326	96.742
StableNormal [53]	8.034	6.568	21.393	43.917	63.740	78.568	93.671	96.785
<b>Ours</b>	<b>6.413</b>	<b>4.897</b>	<b>38.656</b>	<b>56.780</b>	<b>70.938</b>	<b>82.853</b>	<b>95.412</b>	<b>98.063</b>

Table 2. Quantitative comparison on PBR materials estimation and single-image relighting.

Method	Albedo		Roughness		Metallic		Relighting			Latency
	PSNR ↑	RMSE ↓	PSNR ↑	RMSE ↓	PSNR ↑	RMSE ↓	PSNR ↑	SSIM ↑	LPIPS ↓	Average Time ↓
RGB↔X [57]	16.26	0.176	19.21	0.134	16.65	0.199	20.78	0.8927	0.0781	15s
Yi. et al [54]	21.10	0.106	16.88	0.180	20.30	0.144	26.47	0.9316	0.0691	5s
IntrinsicAnything [8]	23.88	0.078	17.25	0.172	22.00	0.134	27.98	0.9474	0.0490	2min
DiLightNet [56]	-	-	-	-	-	-	22.68	0.8751	0.0981	30s
IC-Light [60]	-	-	-	-	-	-	20.29	0.9027	0.0638	1min
<b>Ours</b>	<b>26.62</b>	<b>0.054</b>	<b>23.44</b>	<b>0.085</b>	<b>26.23</b>	<b>0.109</b>	<b>30.12</b>	<b>0.9601</b>	<b>0.0371</b>	5s

However, capturing such pairs in the real world requires specialized photometric equipments and controlled lighting, which is impractical for large-scale collection, while internet images typically lack access to their underlying 3D data, making it infeasible to derive ground-truth surface properties. Therefore, we construct a synthetic dataset *LightProp*, where we curate 80k objects from Objaverse [11], filtering out those of low-quality or without PBR materials.

*LightProp* provides multi-light images and G-buffer maps for every object. Each object is rendered at 5 random views, and for each view, we simulate 5 images under random lighting conditions, including point light, area light, and HDR environment maps. Each view also provides a full set of surface normal and PBR materials, along with multi-light images rendered under known directional lighting. As shown in Fig. 4, we position the camera and point lights on a sphere around the object, where  $\theta$  determines the vertical position of the lights relative to the overhead direction, and  $\varphi$  controls the rotation relative to the camera. In practice, the positions of light sources are fixed during the training of multi-light diffusion model  $g(\cdot)$  and the inference of G-buffer prediction model  $f(\cdot)$ , while randomized light positions are applied for training  $f(\cdot)$  to encourage generalization. More details on dataset construction can be found in the appendix.

## 4. Experiments

We evaluate our method across various tasks. For **normal estimation**, we benchmark against regression-based method DSINE [2] and diffusion-based methods GeoWizard [16], Marigold [25] and StableNormal [53]. For **PBR material prediction**, we compare our method with a data-driven method by Yi et al. [54], an optimization method IntrinsicAnything [8], and a diffusion-based model

RGB↔X [57]. For **image relighting**, we use ground-truth normal maps and predicted PBR materials from baselines [8, 54, 57] to render relit images, serving as relighting baselines. We also compare our method with diffusion-based image relighting models DiLightNet [56] and IC-Light [60], using a captioning model [39] to generate prompts.

### 4.1. Quantitative Evaluation

We calculate metrics on a held-out subset of *LightProp*, consisting of 1,000 randomly selected, unseen objects.

**Normal.** Following prior works [16, 53], we report the comparison results in mean and median angular errors, and accuracy within various angular thresholds. Since we observe promising accuracy within the commonly used thresholds from 5° to 30°, we further report the accuracy under a finer threshold of 3°. As shown in Tab. 1, our method outperforms baselines across all metrics, particularly under finer thresholds, clearly showing the effectiveness.

**Materials and Relighting.** Following previous works, we calculate PSNR and RMSE for albedo, roughness, and metallic maps, and evaluate relit images using PSNR, SSIM, and LPIPS [61]. We also report the average time per frame, calculated by measuring the total time to render 120 relit frames from a single input image and dividing by the number of frames. As shown in Tab. 2, our method shows a clear improvement over baselines. These results demonstrate the effectiveness and efficiency of our approach in predicting accurate material properties and rendering faithful relighting images.

### 4.2. Qualitative Evaluation

We present qualitative comparison results on both the unseen Objaverse subset and in-the-wild images. More visual

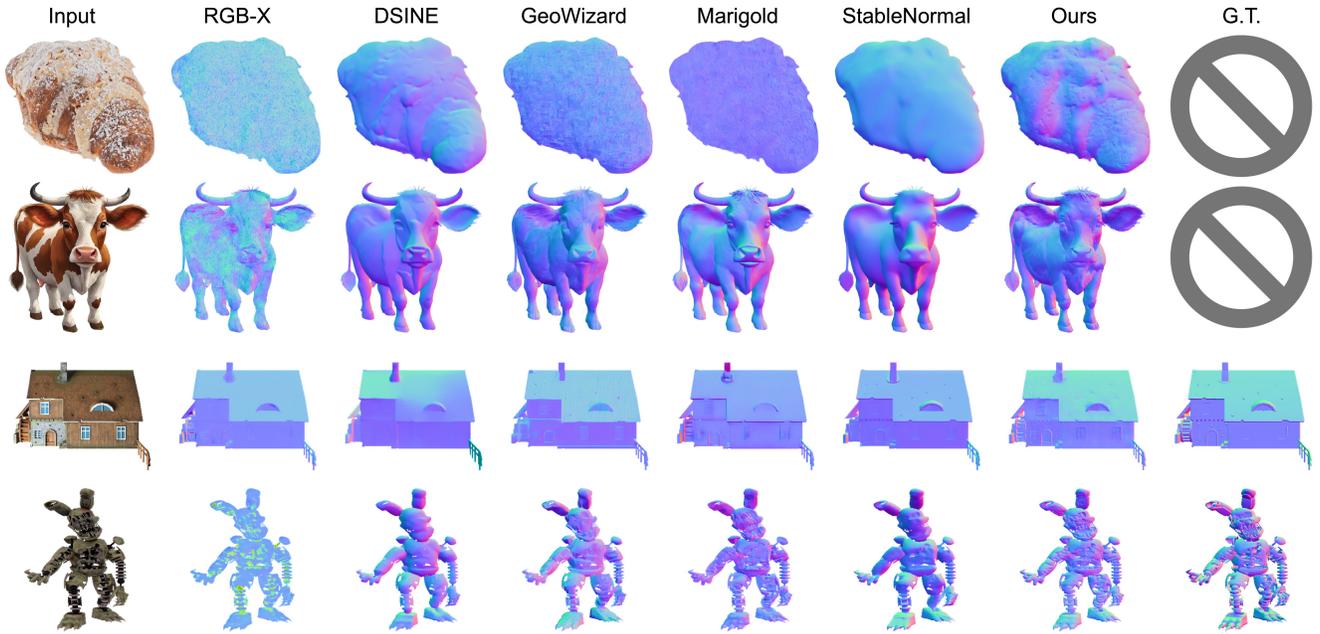


Figure 5. Qualitative comparison on surface normal estimation. Ground truth normals (G.T.) are provided for input images rendered from available 3D objects (the last two rows) and are omitted for in-the-wild images (the first two rows).

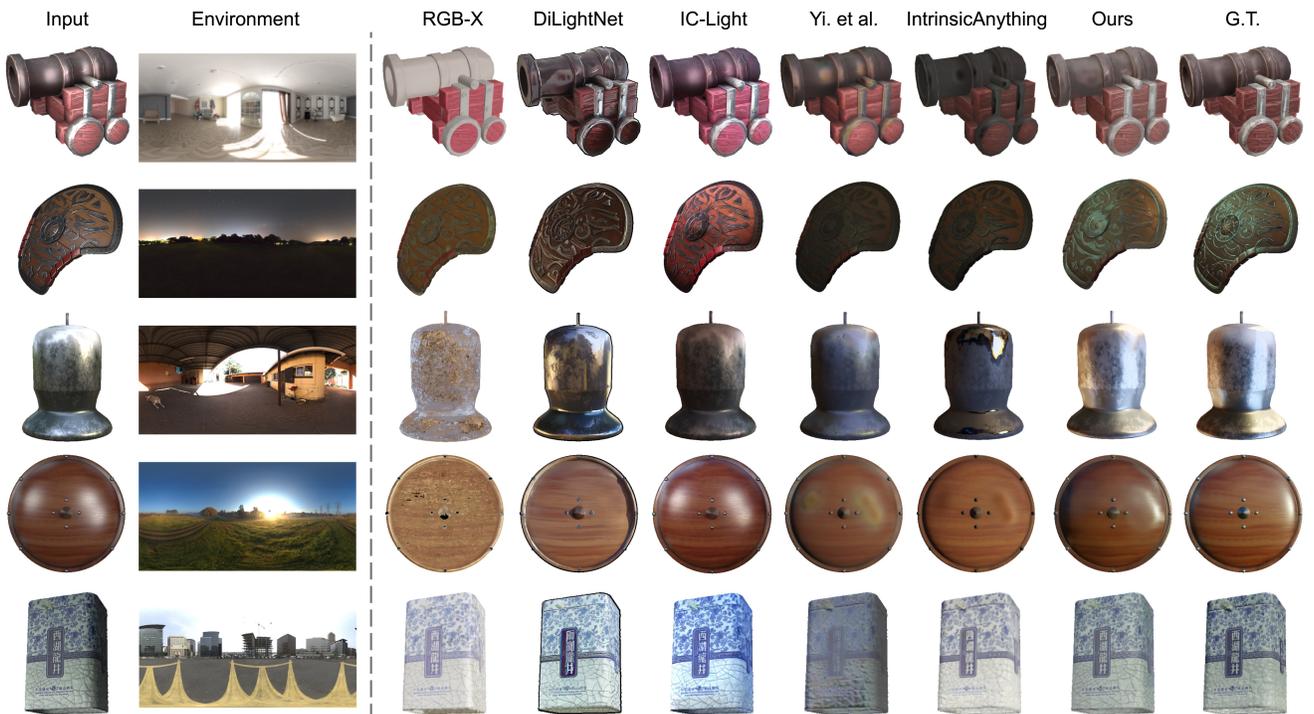


Figure 6. Qualitative comparison on single-image relighting.

results are given in *appendix*.

**Normal.** As shown in Fig. 5, our method produces sharp, coherent normal maps while preserving surface details. For instance, in the cow case, our method accurately captures

the normal variations around the ears. In the robot example, other methods tend to produce over-smoothed or inaccurate normal, while ours demonstrates a clear advantage in capturing complex surface geometries. Please refer to Fig. 16

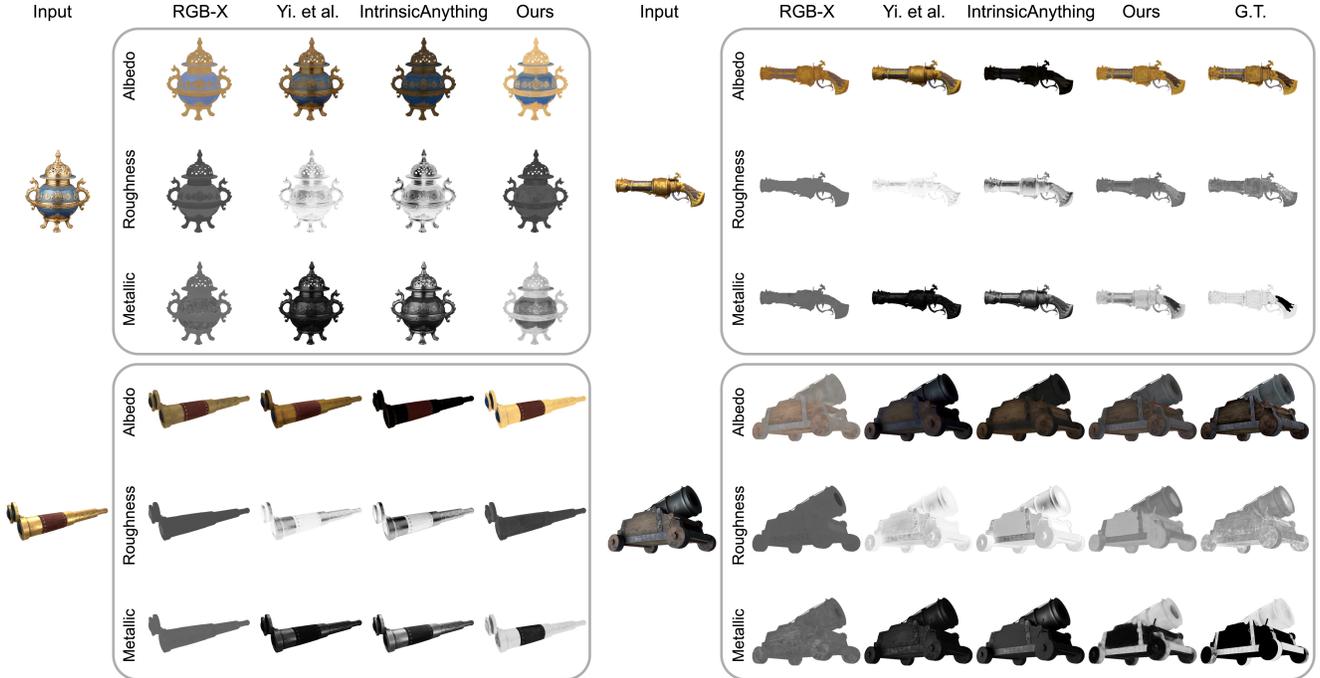


Figure 7. Qualitative comparison on PBR material estimation. Ground truth materials (G.T.) are provided for input images rendered from available 3D objects (the right column) and are omitted for in-the-wild images (the left column).

Table 3. Effects of condition strategies in multi-light diffusion.

	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$
Concatenation	19.32	0.8597	0.0909
Reference Attention	19.87	0.8691	0.0829
Concatenation + Reference Attention	<b>20.01</b>	<b>0.8718</b>	<b>0.0815</b>

for more examples.

**PBR Materials.** As shown in Fig. 7, our approach generates more accurate PBR materials than baselines. Baseline methods fail to remove highlights in their albedo maps, while our approach produces smooth base colors regardless of the illumination conditions of input images. Also, our method is more robust at distinguishing metal and nonmetal materials, while baselines are prone to reflective parts or fail to locate the metallic regions. More examples can be found in Figs. 17 and 18.

**Image Relighting.** As shown in Fig. 6, our approach generates realistic lighting effects and retains details such as Chinese characters in the last example. In contrast, without underlying physical properties, DiLightNet and IC-Light tend to generate over-saturated images, while others are limited in eliminating highlights and shadows from the input image. Video comparisons are provided in our [project page](#). In the appendix, we provide more relighting comparisons in Fig. 19 and more relighting results of our method in Figs. 14 and 15.

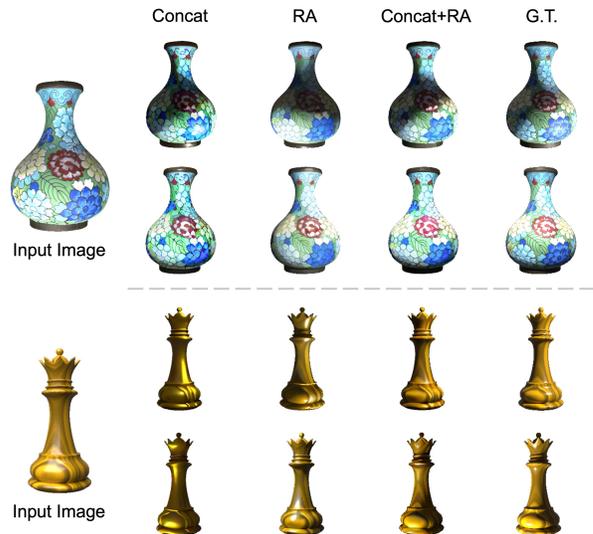


Figure 8. Visualization of different conditioning strategies in multi-light diffusion. *Concat* stands for concatenation. *RA* stands for reference attention.

### 4.3. Ablation Study

Due to the expensive training cost of the full model, we use smaller models for the following ablation experiments.

**Conditioning Strategy for Multi-Light Diffusion.** We ex-

Table 4. Effect of the number of multi-light images on the performance of the large G-buffer model.

Number of Light Images	Albedo		Roughness		Metallic		MAE ↓	5° ↑	Normal		
	PSNR ↑	RMSE ↓	PSNR ↑	RMSE ↓	PSNR ↑	RMSE ↓			7.5° ↑	11.25° ↑	22.5° ↑
0	22.22	0.082	20.99	0.104	18.56	0.136	7.563	45.846	61.425	76.948	95.488
3	23.72	0.068	23.89	0.075	20.66	0.106	4.763	68.344	80.896	89.959	97.928
6	23.82	0.068	24.19	0.072	20.64	0.106	4.275	72.777	83.997	91.730	98.312
9	<b>23.90</b>	<b>0.067</b>	<b>24.36</b>	<b>0.069</b>	<b>20.74</b>	<b>0.105</b>	<b>4.059</b>	<b>74.720</b>	<b>85.092</b>	<b>92.330</b>	<b>98.431</b>

Table 5. Effect of augmentation strategy on the large G-buffer model.

	Albedo		Roughness		Metallic		MAE ↓	5° ↑	Normal		
	PSNR ↑	RMSE ↓	PSNR ↑	RMSE ↓	PSNR ↑	RMSE ↓			7.5° ↑	11.25° ↑	22.5° ↑
w/o augmentation	21.69	0.087	20.46	0.110	16.61	0.179	7.080	52.235	67.032	80.115	94.802
w/ augmentation	<b>22.36</b>	<b>0.081</b>	<b>21.39</b>	<b>0.099</b>	<b>18.81</b>	<b>0.135</b>	<b>6.342</b>	<b>55.893</b>	<b>70.326</b>	<b>82.848</b>	<b>96.230</b>

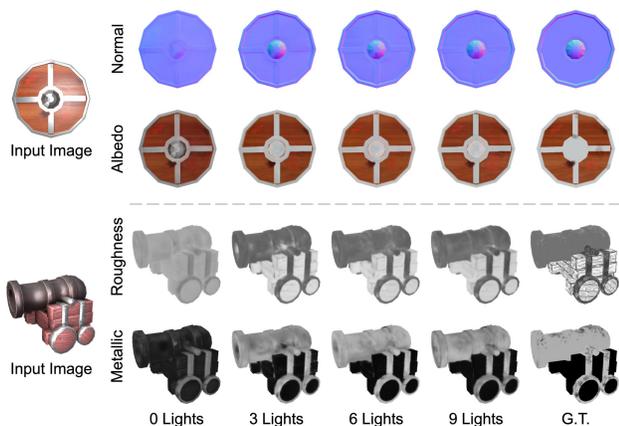


Figure 9. Visualization of using different numbers of multi-light images. We evaluate the G-Buffer prediction model with different numbers of novel-light images (0, 3, 6, and 9) as conditions.

explore three different settings, concatenation, reference attention (*RA*), and our hybrid approach. The quantitative analyses are given in Tab. 3. As shown in Fig. 8, while *Concat* captures correct highlights and shadows, it often results in over-saturated colors or inaccurately rendered surface textures, as seen in the excessive brightness on the vase and inconsistent color tones on the chess piece. *RA*, on the other hand, fails to reflect faithful lighting effects. In contrast, the hybrid approach yields the best qualitative and quantitative performances.

**Number of Multi-Light Images for Prediction.** To examine how multi-light images affect performance, we evaluate the large G-buffer model with varying numbers of rendered light images (0, 3, 6, and 9). As shown in Tab. 4, the performances improve sharply from 0 to 3 images by reducing ambiguity, and steadily improve with more provided images. The same conclusion is also observed in Fig. 9, where leveraging multi-light images yields sharper normal and better PBR maps.

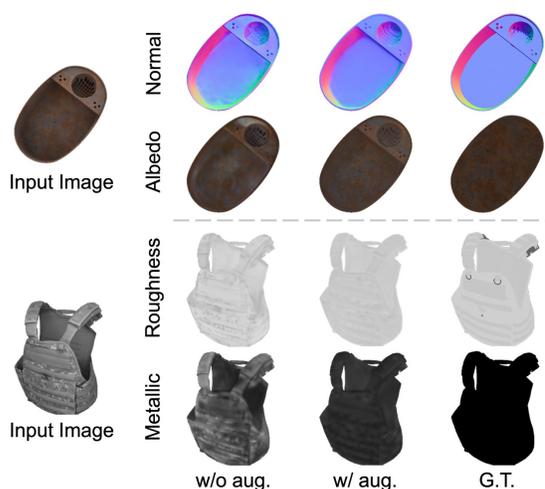


Figure 10. Visualization of the augmentation strategy.

**Effects of Augmentation Strategy.** We examine the impact of data augmentation on enhancing the robustness and generalization of the G-buffer prediction model. As shown in Tab. 5 and Fig. 10, the proposed augmentation strategy improves the model’s ability to produce consistent and accurate outputs, demonstrating increased invariance to artifacts introduced by the multi-light diffusion model. This augmentation effectively bridges the gap caused by noise, color inconsistencies, and other disturbances.

## 5. Conclusion

In this work, we present *Neural LightRig*, a framework capable of estimating accurate surface normals and PBR materials from a single image. Leveraging a multi-light diffusion model, we generated consistent relit images under various directional light sources. These generated images significantly reduce the inherent ambiguity when estimating surface properties, serving as enriched conditions for the G-Buffer prediction model. Extensive experiments demon-

strate that our method achieves significant improvements in both quality and generalizability. Future work will focus on extending this approach to more complex scenes and integrating it with 3D reconstruction systems.

## References

- [1] Stability AI. Stable diffusion v2.1. <https://huggingface.co/stabilityai/stable-diffusion-2-1>, 2023. 4
- [2] Gwangbin Bae and Andrew J. Davison. Rethinking inductive biases for surface normal estimation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2, 5
- [3] Gwangbin Bae, Ignas Budvytis, and Roberto Cipolla. Estimating and exploiting the aleatoric uncertainty in surface normal estimation. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, page 13117–13126. IEEE, 2021. 2
- [4] Aayush Bansal, Bryan Russell, and Abhinav Gupta. Marr revisited: 2d-3d alignment via surface normal prediction. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, page 5965–5974. IEEE, 2016. 2
- [5] Manel Baradad, Yuanzhen Li, Forrester Cole, Michael Rubinstein, Antonio Torralba, William T. Freeman, and Varun Jampani. Background prompting for improved object depth, 2023. 2
- [6] Jonathan T Barron and Jitendra Malik. Shape, albedo, and illumination from a single image of an unknown object. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 334–341. IEEE, 2012. 2
- [7] Mark Boss, Raphael Braun, Varun Jampani, Jonathan T. Barron, Ce Liu, and Hendrik P.A. Lensch. NerD: Neural reflectance decomposition from image collections. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, 2021. 3
- [8] Xi Chen, Sida Peng, Dongchen Yang, Yuan Liu, Bowen Pan, Chengfei Lv, and Xiaowei Zhou. Intrinsicanything: Learning diffusion priors for inverse rendering under unknown illumination, 2024. 3, 5
- [9] Blender Online Community. *Blender - a 3D modelling and rendering package*. Blender Foundation, Stichting Blender Foundation, Amsterdam, 2018. 2, 12
- [10] Paul Debevec, Tim Hawkins, Chris Tchou, Haarm-Pieter Duiker, Westley Sarokin, and Mark Sagar. Acquiring the reflectance field of a human face. In *Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques*, page 145–156, USA, 2000. ACM Press/Addison-Wesley Publishing Co. 2, 3
- [11] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13142–13153, 2023. 2, 3, 5, 12
- [12] Tien Do, Khiem Vuong, Stergios I. Roumeliotis, and Hyun Soo Park. Surface normal estimation of tilted images via spatial rectifier. In *Proc. of the European Conference on Computer Vision*, Virtual Conference, 2020. 2
- [13] O. Drbohlav and M. Chaniler. Can two specular pixels calibrate photometric stereo? In *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, pages 1850–1857 Vol. 2, 2005. 2, 3
- [14] Ainaz Eftekhari, Alexander Sax, Jitendra Malik, and Amir Zamir. Omnidata: A scalable pipeline for making multi-task mid-level vision datasets from 3d scans. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, page 10766–10776. IEEE, 2021. 2
- [15] David F. Fouhey, Abhinav Gupta, and Martial Hebert. Data-driven 3d primitives for single image understanding. In *2013 IEEE International Conference on Computer Vision*, pages 3392–3399, 2013. 2
- [16] Xiao Fu, Wei Yin, Mu Hu, Kaixuan Wang, Yuexin Ma, Ping Tan, Shaojie Shen, Dahua Lin, and Xiaoxiao Long. Geowizard: Unleashing the diffusion priors for 3d geometry estimation from a single image. In *ECCV*, 2024. 2, 5
- [17] Jon Hasselgren, Nikolai Hofmann, and Jacob Munkberg. Shape, light, and material decomposition from images using monte carlo rendering and denoising. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, 2024. Curran Associates Inc. 2, 3
- [18] Zexin He and Tengfei Wang. Openlrm: Open-source large reconstruction models. <https://github.com/3DTopia/OpenLRM>, 2023. 2
- [19] Derek Hoiem, Alexei A. Efros, and Martial Hebert. Automatic photo pop-up. *ACM Trans. Graph.*, 24(3):577–584, 2005. 2
- [20] Derek Hoiem, Alexei A. Efros, and Martial Hebert. Recovering surface layout from an image. *International Journal of Computer Vision: Special Issue on Celebrating Kanade's Vision*, 75(1):151 – 172, 2007. 2
- [21] Yicong Hong, Kai Zhang, Jiuxiang Gu, Sai Bi, Yang Zhou, Difan Liu, Feng Liu, Kalyan Sunkavalli, Trung Bui, and Hao Tan. LRM: Large reconstruction model for single image to 3d. In *The Twelfth International Conference on Learning Representations*, 2024. 2
- [22] Xin Huang, Tengfei Wang, Ziwei Liu, and Qing Wang. Material anything: Generating materials for any 3d object via diffusion. *arXiv*, 2024. 3
- [23] 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 *Advances in Neural Information Processing Systems*, 2024. 2, 3
- [24] Berk Kaya, Suryansh Kumar, Carlos Oliveira, Vittorio Ferrari, and Luc Van Gool. Multi-view photometric stereo revisited. In *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, page 3125–3134. IEEE, 2023. 3
- [25] Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2, 5

- [26] L’ubor Ladický, Bernhard Zeisl, and Marc Pollefeys. Discriminatively trained dense surface normal estimation. In *ECCV*, pages 468–484. Springer International Publishing, 2014. 2
- [27] Samuli Laine, Janne Hellsten, Tero Karras, Yeongho Seol, Jaakko Lehtinen, and Timo Aila. Modular primitives for high-performance differentiable rendering. *ACM Transactions on Graphics*, 39(6), 2020. 2
- [28] Marc Levoy, Kari Pulli, Brian Curless, Szymon Rusinkiewicz, David Koller, Lucas Pereira, Matt Gintzton, Sean Anderson, James Davis, Jeremy Ginsberg, Jonathan Shade, and Duane Fulk. The digital michelangelo project: 3d scanning of large statues. In *Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques*, page 131–144, USA, 2000. ACM Press/Addison-Wesley Publishing Co. 2, 3
- [29] Bo Li, Chunhua Shen, Yuchao Dai, Anton van den Hengel, and Mingyi He. Depth and surface normal estimation from monocular images using regression on deep features and hierarchical crfs. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1119–1127, 2015. 2
- [30] Jiahao Li, Hao Tan, Kai Zhang, Zexiang Xu, Fujun Luan, Yinghao Xu, Yicong Hong, Kalyan Sunkavalli, Greg Shakhnarovich, and Sai Bi. Instant3d: Fast text-to-3d with sparse-view generation and large reconstruction model. In *The Twelfth International Conference on Learning Representations*, 2024. 2, 13
- [31] Daniel Lichy, Jiaye Wu, Soumyadip Sengupta, and David W. Jacobs. Shape and material capture at home. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, page 6119–6129. IEEE, 2021. 3
- [32] 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 *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, page 10072–10083. IEEE, 2024. 2
- [33] 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. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, page 9264–9275. IEEE, 2023. 2
- [34] Yunfei Liu, Yu Li, Shaodi You, and Feng Lu. Unsupervised learning for intrinsic image decomposition from a single image. In *CVPR*, 2020. 2, 3
- [35] 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. In *The Twelfth International Conference on Learning Representations*, 2024. 2
- [36] Linjie Lyu, Ayush Tewari, Marc Habermann, Shunsuke Saito, Michael Zollhöfer, Thomas Leimkühler, and Christian Theobalt. Diffusion posterior illumination for ambiguity-aware inverse rendering. *ACM Transactions on Graphics*, 42(6), 2023. 3
- [37] Xiaojuan Qi, Zhengzhe Liu, Renjie Liao, Philip H. S. Torr, Raquel Urtasun, and Jiaya Jia. Geonet++: Iterative geometric neural network with edge-aware refinement for joint depth and surface normal estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(2):969–984, 2022. 2
- [38] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, 2022. 2
- [39] Salesforce. Blip-2, opt-2.7b, pre-trained only. <https://huggingface.co/Salesforce/blip2-opt-2.7b>, 2023. 5
- [40] Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. In *International Conference on Learning Representations*, 2022. 4
- [41] Shen Sang and M. Chandraker. Single-shot neural relighting and svbrdf estimation. In *ECCV*, 2020. 3
- [42] Jian Shi, Yue Dong, Hao Su, and Stella X. Yu. Learning non-lambertian object intrinsics across shapenet categories. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5844–5853, 2017. 3
- [43] 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, 2023. 2, 3
- [44] Yichun Shi, Peng Wang, Jianglong Ye, Long Mai, Kejie Li, and Xiao Yang. MVDream: Multi-view diffusion for 3d generation. In *The Twelfth International Conference on Learning Representations*, 2024. 2
- [45] 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 *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, page 7491–7500. IEEE, 2021. 3
- [46] Jiayang Tang, Zhaoxi Chen, Xiaokang Chen, Tengfei Wang, Gang Zeng, and Ziwei Liu. Lgm: Large multi-view gaussian model for high-resolution 3d content creation. In *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part IV*, page 1–18, Berlin, Heidelberg, 2024. Springer-Verlag. 2
- [47] Giuseppe Vecchio and Valentin Deschaintre. Matsynth: A modern pbr materials dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 3
- [48] Rui Wang, David Geraghty, Kevin Matzen, Richard Szeliski, and Jan-Michael Frahm. Vplnet: Deep single view normal estimation with vanishing points and lines. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 686–695, 2020. 2
- [49] Tengfei Wang, Ting Zhang, Bo Zhang, Hao Ouyang, Dong Chen, Qifeng Chen, and Fang Wen. Pretraining is all you need for image-to-image translation. In *arXiv*, 2022. 2
- [50] Zhenwei Wang, Tengfei Wang, Zexin He, Gerhard Hancke, Ziwei Liu, and Rynson WH Lau. Phidias: A generative

- model for creating 3d content from text, image, and 3d conditions with reference-augmented diffusion. *arXiv preprint arXiv:2409.11406*, 2024. 2
- [51] Robert J. Woodham. *Photometric method for determining surface orientation from multiple images*, page 513–531. MIT Press, Cambridge, MA, USA, 1989. 2, 3
- [52] Guangkai Xu, Yongtao Ge, Mingyu Liu, Chengxiang Fan, Kangyang Xie, Zhiyue Zhao, Hao Chen, and Chunhua Shen. What matters when repurposing diffusion models for general dense perception tasks?, 2024. 2
- [53] Chongjie Ye, Lingteng Qiu, Xiaodong Gu, Qi Zuo, Yushuang Wu, Zilong Dong, Liefeng Bo, Yuliang Xiu, and Xiaoguang Han. Stablenormal: Reducing diffusion variance for stable and sharp normal. *ACM Transactions on Graphics (TOG)*, 2024. 2, 5
- [54] Renjiao Yi, Chenyang Zhu, and Kai Xu. Weakly-supervised single-view image relighting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8402–8411, 2023. 2, 3, 5
- [55] Ye Yu and William A. P. Smith. Inverserendernet: Learning single image inverse rendering. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, page 3150–3159. IEEE, 2019. 3
- [56] Chong Zeng, Yue Dong, Pieter Peers, Youkang Kong, Hongzhi Wu, and Xin Tong. Dilightnet: Fine-grained lighting control for diffusion-based image generation. In *ACM SIGGRAPH 2024 Conference Papers*, 2024. 2, 3, 5
- [57] Zheng Zeng, Valentin Deschaintre, Iliyan Georgiev, Yannick Hold-Geoffroy, Yiwei Hu, Fujun Luan, Ling-Qi Yan, and Miloš Hašan. Rgb $\leftrightarrow$ x: Image decomposition and synthesis using material- and lighting-aware diffusion models. In *ACM SIGGRAPH 2024 Conference Papers*, New York, NY, USA, 2024. Association for Computing Machinery. 2, 3, 5
- [58] Jingyang Zhang, Yao Yao, Shiwei Li, Jingbo Liu, Tian Fang, David McKinnon, Yanghai Tsin, and Long Quan. Neilf++: Inter-reflectable light fields for geometry and material estimation. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, 2023. 3
- [59] Lyumin Zhang. Reference-only control. <https://github.com/Mikubill/sd-webui-controlnet/discussions/1236>, 2023. 3
- [60] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Ic-light github page, 2024. 5
- [61] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 5
- [62] Zhenyu Zhang, Zhen Cui, Chunyan Xu, Yan Yan, Nicu Sebe, and Jian Yang. Pattern-affinitive propagation across depth, surface normal and semantic segmentation. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, page 4101–4110. IEEE, 2019. 2

# Appendix

## A. Dataset Details

In the main paper, we provided an overview of the *LightProp* dataset, designed specifically to address the challenges of learning robust multi-light image generation and geometry-material estimation. Here, we detail the data curation and rendering configurations.

### A.1. Data Curation

Objaverse [11] originally contains around 800,000 synthetic objects across various categories and styles. To ensure high-quality content for *LightProp*, we implemented a rigorous curation process. First, we filtered out objects with extreme thinness or unbalanced proportions, such as objects with large surface areas but minimal thickness or depth, which often distort lighting interactions and hinder effective learning. Additionally, we excluded objects that originated from 3D scans or those representing entire scenes, as these typically contain irrelevant environmental details that are less suitable for our framework. Finally, objects lacking essential PBR material maps (albedo, roughness, and metallic maps) were removed to ensure comprehensive material data for training. This selection process resulted in a refined subset of around 80,000 high-quality objects for *LightProp*.

### A.2. Rendering Setup

The *LightProp* dataset is created using the Cycles rendering engine in Blender [9], with each image generated at 128 samples per pixel and accelerated using CUDA. To introduce diversity in object orientation and perspective, each object is rendered from five distinct viewpoints: a front view, a right view, a top view, and two random views sampled on a surrounding sphere. For each viewpoint, we apply five distinct lighting conditions, comprising a point light, an area light, and three HDR environment maps randomly selected from 25 high-quality maps. To set up our directional lighting, we position eight lights around the camera and place one additional light directly at the camera’s position. The lighting orientations are parameterized by spherical coordinates  $\theta$  and  $\varphi$ , specifically configured as:

$$\theta_i = i \cdot \frac{\pi}{4} \quad \text{for } i = 0, 1, \dots, 8, \quad (6)$$

$$\varphi_i = \{1, 2, 1, 2, 1, 2, 1, 2, 0\} \cdot \frac{\pi}{6}. \quad (7)$$

This arrangement ensures diverse lighting directions to enhance shading and reflectance variations in multi-light images, which are essential for accurate geometry and material estimation. In addition to the multi-light images, each object view is paired with ground-truth G-buffer maps, including surface normals, albedo, roughness, and metallic maps. These G-buffers, rendered via Blender’s physically-based

pipeline, provide the necessary supervision for training in surface normal and PBR material prediction.

## B. Implementation Details

### B.1. Multi-Light Diffusion

We build our multi-light diffusion model on top of Stable Diffusion v2-1. As discussed in the main paper, we adopt a two-phase training scheme to adapt this pre-trained model for multi-light image generation. In the initial phase, we tune the first convolution layer, all parameters in the self-attention layers, and only the key and value parameters in the cross-attention layers. This phase runs for 80,000 steps with a peak learning rate of  $1 \times 10^{-4}$  and a total batch size of 128, following a cosine annealing schedule with 2,000 warm-up steps. We use the AdamW optimizer with  $\beta_1 = 0.9$ ,  $\beta_2 = 0.999$ , and a weight decay of 0.01, and enable bf16 mixed precision to accelerate the training. Additionally, we apply gradient clipping with a maximum norm of 1.0 to stabilize training and incorporate classifier-free guidance, with a probability of dropping the conditioning set to 0.1. In the following phase, we further fine-tune the full model for another 80,000 steps at a significantly lower peak learning rate of  $5 \times 10^{-6}$  with the same training particulars. Both of the two phases are trained with an input image resolution of  $256 \times 256$ , and a multi-light output of  $768 \times 768$ . In total, the complete training process of our multi-light diffusion model takes approximately 2.5 days on 32 NVIDIA A100 (80G) GPUs.

### B.2. Large G-Buffer Prediction Model

**Architecture.** Our large G-buffer prediction model takes as input a single image with 4 channels (including alpha), combined with multi-light images comprising 9 lighting conditions, each with 3 channels, resulting in a total of  $4+9 \times 3 = 31$  input channels. The output consists of 8 channels, representing the surface normals, albedo, roughness, and metallic maps (3, 3, 1, and 1 channel, respectively). The regression U-Net architecture comprises four down-sampling blocks with progressively increasing channels of 224, 448, 672, and 896, followed by a bottleneck block with 896 channels, and then four up-sampling blocks with correspondingly decreasing channels of 896, 672, 448, and 224. Each block contains two residual layers with Group Normalization (using 32 groups), and SiLU activation. Attention mechanisms, implemented in a pre-norm style, are applied in all but the first down-sampling block and the last up-sampling block, using an attention head dimension of 8. Within each block, up-sampling and down-sampling are performed via a convolutional layer placed after the two residual layers. To encode the spherical coordinates  $\{\theta^i, \varphi^i\}$  associated with each lighting condition, we employ

<https://huggingface.co/stabilityai/stable-diffusion-2-1>

sinusoidal embeddings. Each scalar  $\theta$  or  $\varphi$  is projected to a higher dimension of  $d_{scalar} = 224$  and we concatenate these projected vectors into a single  $9 \times 2 \times 224 = 4032$  dimensional vector, which is subsequently embedded by a 2-layer MLP, producing an illumination embedding with a final dimensionality of  $d_{emb} = 896$ . This embedding is modulated to each block in the U-Net with adaptive group normalization. For the smaller models in our ablation study, we use a U-Net with down-sampling blocks at 128, 256, 384, and 512 channels, mirrored in the up-sampling blocks, along with a 512-channel bottleneck block.

**Training Details.** We apply weighted loss contributions to balance  $\mathcal{L}_{normal}$  and  $\mathcal{L}_{PBR}$ . Specifically, we set a 4 : 1 ratio for surface normals relative to PBR materials. Additionally, we apply a stabilization factor of  $\lambda_1 = 0.25$  for the MSE term in  $\mathcal{L}_{normal}$ , as outlined in the main paper. Given the computational demands of high-resolution feature maps, especially with attention layers, we employ a two-phase training strategy, gradually transitioning from low to high resolutions. In the initial phase, we train at a resolution of  $256 \times 256$  to establish core feature representations, running for 60,000 steps with a batch size of 128. This phase includes 1,500 warm-up steps, a peak learning rate of  $1 \times 10^{-4}$ , and a weight decay of 0.01, using a cosine annealing schedule and the AdamW optimizer with  $\beta_1 = 0.9$  and  $\beta_2 = 0.999$ . Training on 32 NVIDIA A100 (80G) GPUs, this phase completes in approximately 20 hours. Following this foundational phase, we move to a higher resolution of  $512 \times 512$ , allowing the model to capture finer details essential for precise geometry and material predictions. This fine-tuning phase involves a reduced learning rate of  $2 \times 10^{-5}$  and runs for an additional 30,000 steps on the same setup of 32 NVIDIA A100 (80G) GPUs, completing in approximately 7 days. All other training parameters are kept consistent with the initial phase.

**Augmentation Details.** In the main paper, we introduced the augmentations to bridge the gap between our multi-light diffusion model and the large G-buffer prediction model. For *Random Degradation*, we down-sample each multi-light image to a lower resolution uniformly sampled from  $\mathcal{U}(128, 256)$  and then up-sample it back to the original resolution of 256. Following this, we apply grid distortion with a perturbation strength sampled from  $\mathcal{U}(0.15, 0.3)$  to simulate geometrical misalignments. For *Random Intensity*, we convert the multi-light images to HSV format and adjust the brightness channel using an image-level scaling factor from  $\mathcal{U}(0.9, 1.3)$ . Additionally, we apply pixel-level noise by scaling each pixel independently with a factor sampled from  $\mathcal{N}(1, 0.05)$ . The input image receives a separate brightness adjustment factor sampled from  $\mathcal{U}(0.9, 1.1)$ . For *Random Orientation*, all spherical coordinates are perturbed by an angular gaussian noise in radians.  $\theta^i$  receive a noise sampled from  $\mathcal{N}(0, 0.1)$  and are wrapped with modulus  $2\pi$ .



Figure 11. Failure case.

$\varphi^i$  are perturbed with noise from  $\mathcal{N}(0, 0.02)$  and clamped within  $[0, \frac{\pi}{2}]$ . The above three augmentations are triggered independently with a probability of 0.6. For *Data Mixing*, this augmentation is applied with a probability of 0.3. We generate multi-light images from our diffusion model with a classifier-free guidance scale of 2.0 over 75 inference steps. Additionally, inspired by prior work on multi-view reconstruction [30], we shuffle the order of the multi-light images during training with a probability of 0.5 to encourage robustness in learning features across varied lighting sequences, thereby reducing dependency on any specific lighting arrangement.

## C. Limitations

While our approach demonstrates strong performance, several limitations remain. First, for input images with extreme highlights or shadow areas, our method struggles to fully remove illumination effects in the predicted albedo maps, as shown in Fig. 11. Additionally, the resolution of the backbone multi-light diffusion model ( $256 \times 256$ ) limits the level of detail achievable in the generated multi-light images, subsequently constraining the final normal and material predictions. Increasing the model’s resolution could enhance the quality of the predicted surface properties. Finally, our method is currently designed for objects rather than full scenes, limiting its applicability in complex, multi-object environments.

## D. Additional Results

### D.1. Our Results

Figs. 12 and 13 present examples of our full pipeline output, including input images, generated multi-light images, estimated surface normals, PBR materials, and relit images under various environment maps. These results showcase the robustness of our approach in generating consistent geometry and material estimates and realistic relighting effects across different lighting conditions. Additionally, Figs. 14 and 15 showcase extended single-image relighting results of our method under an even broader range of environment maps, further highlighting the model’s ability to generate high-quality, adaptable relit images across diverse lighting setups. These results illustrate the robustness in managing various lighting conditions and further demonstrate the efficacy of our approach.

## D.2. Comparison Results

In Fig. 16, Fig. 17, Fig. 18, and Fig. 19 we offer more comparison results for surface normal estimation, PBR material estimation, and single-image relighting. These comparisons further demonstrate the advantages of our method over baseline approaches in accurately capturing surface details, material properties, and producing realistic relit images under diverse lighting conditions.

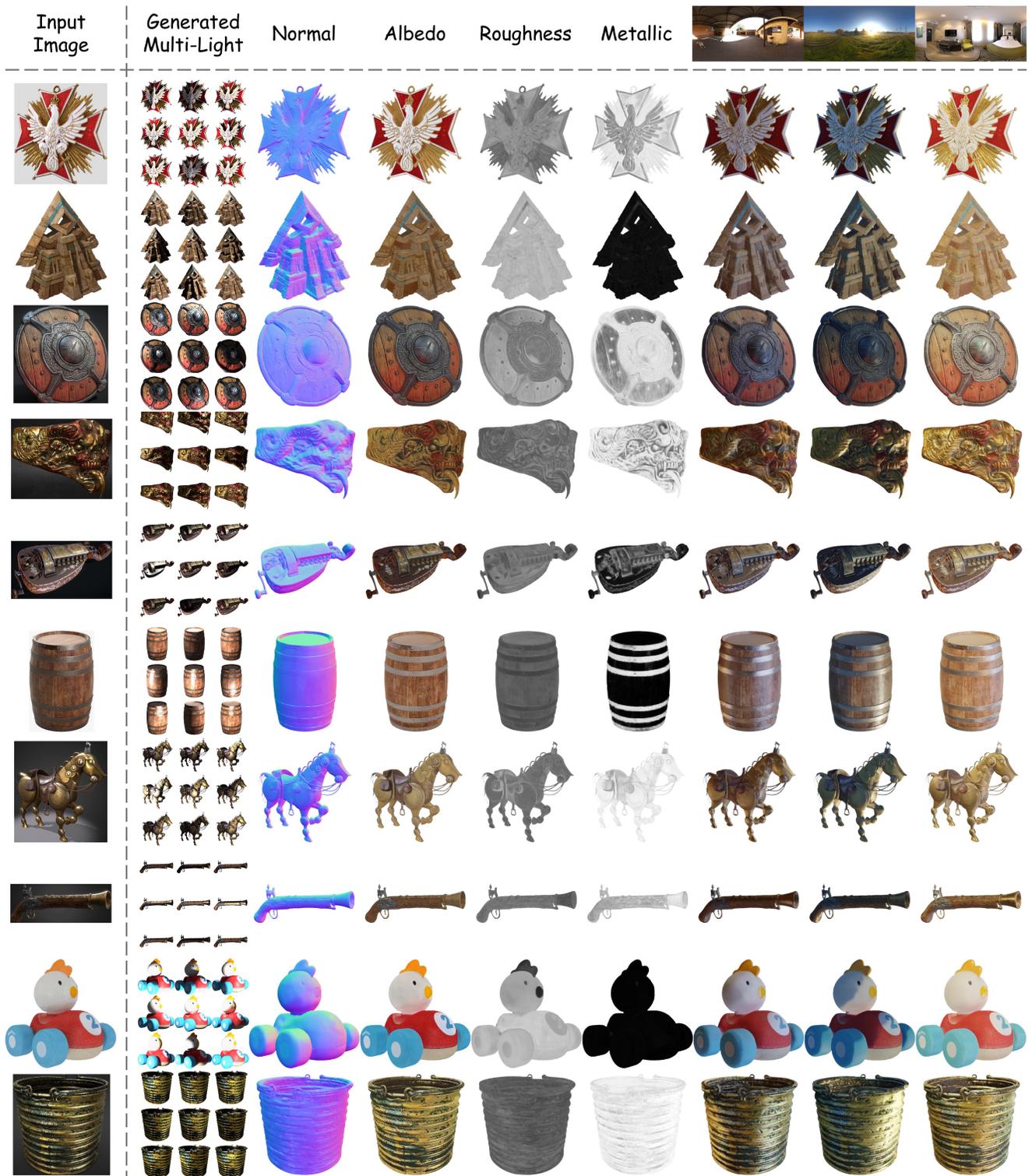


Figure 12. More results of our method.



Figure 13. More results of our method.

Input  
Image



Figure 14. More single-image relighting results of our method.



Figure 15. More single-image relighting results of our method.



Figure 16. More comparisons on surface normal estimation.

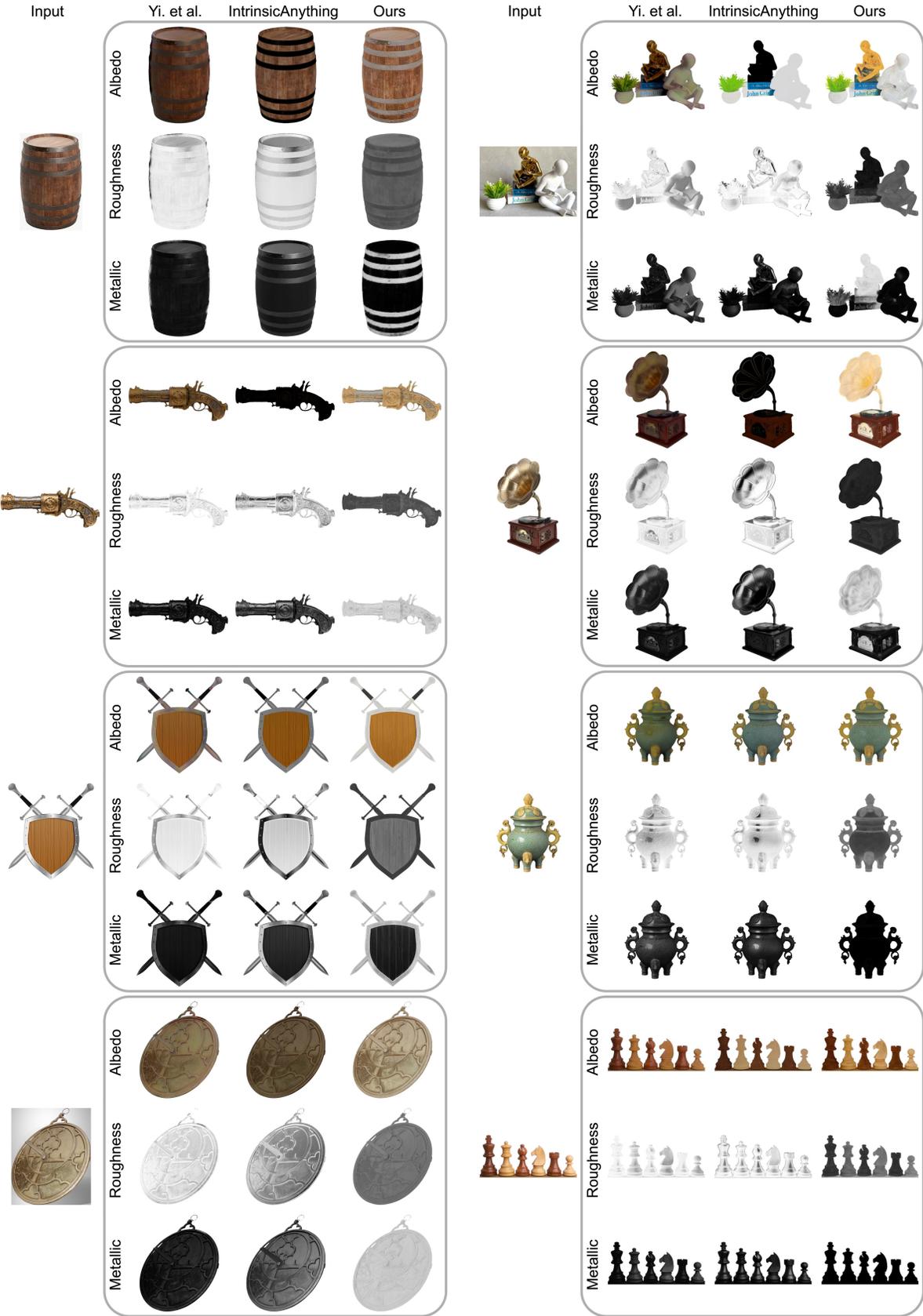


Figure 17. More comparisons on PBR material estimation.



Figure 18. More comparisons on PBR material estimation.



Figure 19. More comparisons on single-image relighting.