

# Diffusion-Guided Relighting for Single-Image SVBRDF Estimation

YOUXIN XING, School of Software, Shandong University, China

ZHENG ZENG, University of California, Santa Barbara, NVIDIA, USA

YOUYANG DU, School of Software, Shandong University, China

LU WANG\*, School of Software, Shandong University, China

BEIBEI WANG\*, School of Intelligence Science and Technology, Nanjing University, China

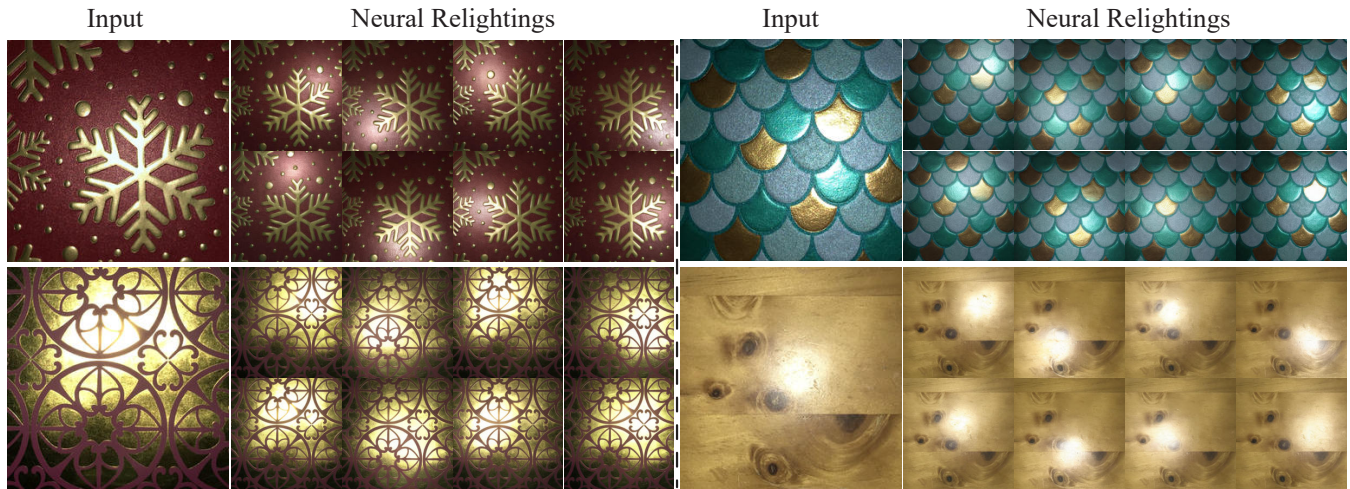


Fig. 1. Using a diffusion model guided by a shuffle-based background consistency module and a specular prior reuse strategy, we generate relighting-consistent and highlight-stable neural relighting materials from a single real-world material photograph.

Recovering high-fidelity spatially varying bidirectional reflectance distribution function (SVBRDF) maps from a single image remains an ill-posed and challenging problem, especially in the presence of saturated highlights. Existing methods often fail to reconstruct the underlying texture in regions overwhelmed by intense specular reflections. This kind of bake-in artifacts caused by highlight corruption can be greatly alleviated by providing a series of material images under different lighting conditions. To this end, our key insight is to leverage the strong priors of diffusion models to generate images of the same material under varying lighting conditions. These generated images are then used to aid a multi-image SVBRDF estimator in recovering highlight-free reflectance maps. However, strong highlights in the input image lead to inconsistencies across the relighting results. Moreover, texture reconstruction becomes unstable in saturated regions, with variations

in background structure, specular shape, and overall material color. These artifacts degrade the quality of SVBRDF recovery. To address this issue, we propose a shuffle-based background consistency module that extracts stable background features and implicitly identifies saturated regions. This guides the diffusion model to generate coherent content while preserving material structures and details. Furthermore, to stabilize the appearance of generated highlights, we introduce a lightweight specular prior encoder that estimates highlight features and then performs grid-based latent feature translation, injecting consistent specular contour priors while preserving material color fidelity. Both quantitative analysis and qualitative visualization demonstrate that our method enables stable neural relighting from a single image and can be seamlessly integrated into multi-input SVBRDF networks to estimate highlight-free reflectance maps.

\*Corresponding authors

Authors' Contact Information: Youxin Xing, School of Software, Shandong University, Jinan, Shandong, China, youxinxing@mail.sdu.edu.cn; Zheng Zeng, University of California, Santa Barbara, NVIDIA, Santa Barbara, California, USA, zhengzeng@ucsb.edu; Youyang Du, School of Software, Shandong University, Jinan, Shandong, China, duyuyang957@gmail.com; Lu Wang, School of Software, Shandong University, Jinan, Shandong, China, luwang\_hcivr@sdu.edu.cn; Beibei Wang, School of Intelligence Science and Technology, Nanjing University, Suzhou, Jiangsu, China, beibei.wang@nju.edu.cn.

SA Conference Papers '25, Hong Kong, Hong Kong

© 2025 Copyright held by the owner/author(s). Publication rights licensed to ACM. This is the author's version of the work. It is posted here for your personal use. Not for redistribution. The definitive Version of Record was published in *SIGGRAPH Asia 2025 Conference Papers (SA Conference Papers '25)*, December 15–18, 2025, Hong Kong, Hong Kong, <https://doi.org/10.1145/3757377.3763809>.

CCS Concepts: • **Computing methodologies** → **Rendering; Reflectance modeling.**

Additional Key Words and Phrases: Neural Relighting, SVBRDF, Capture

**ACM Reference Format:**

Youxin Xing, Zheng Zeng, Youyang Du, Lu Wang, and Beibei Wang. 2025. Diffusion-Guided Relighting for Single-Image SVBRDF Estimation. In *SIGGRAPH Asia 2025 Conference Papers (SA Conference Papers '25)*, December 15–18, 2025, Hong Kong, Hong Kong. ACM, New York, NY, USA, 10 pages. <https://doi.org/10.1145/3757377.3763809>

## 1 Introduction

Estimating spatially varying bidirectional reflectance distribution function (SVBRDF) maps from images is a core problem in appearance modeling. Such maps enable realistic rendering, material editing, and content creation, but manual design or controlled capture remains costly in terms of time and labor.

To reduce acquisition cost, recent deep learning methods recover SVBRDFs from images captured with a phone flash. Multi-image approaches [Deschaintre et al. 2019; Gao et al. 2019; Guo et al. 2020; Luo et al. 2024b] exploit viewpoint and lighting variations to enlarge receptive fields and mitigate saturated highlights via redundancy, yielding robustness under strong specularities. However, their reliance on calibrated multi-view or multi-light capture increases acquisition complexity and reduces practicality.

Single-image methods [Guo et al. 2023; Luo et al. 2024a; Zhou and Kalantari 2022] further ease capture but often fail under strong highlights, where sensor saturation hides diffuse textures. Consequently, predicted SVBRDF maps contain baked-in patterns and yield artifacts under novel illumination.

Early work by Henzler et al. [2021] exploited material stationarity, using self-similarity to inpaint highlight-corrupted regions and partially alleviate bake-in artifacts. Guo et al. [2021] instead proposed a highlight-aware dual-stream network, but the limited generative capacity of convolutional neural networks prevents hallucinating plausible structure in large saturated areas.

In this paper, we explore a way that integrates single-image input to recover clean SVBRDF maps. Our key insight is to synthesize diverse relighting observations of the same material by leveraging the strong priors of a diffusion model [Rombach et al. 2022; Zhang et al. 2023]. These relighting images provide rich, complementary supervision for multi-image SVBRDF estimators. However, due to the inherent stochasticity of diffusion models, the different generated material images always suffer from structural inconsistencies, especially the texture fluctuation in the original highlight regions.

To address this, we condition the diffusion model on disentangled material features, separating diffuse (background) and specular (highlight) components. We further introduce a shuffle-based background consistency module that learns stable background features while implicitly marking saturated regions, guiding the model to generate coherent textures in highlight-dominated areas.

As background features enhance structural consistency, we introduce a specular prior encoder that extracts implicit highlight features from the input and translates them according to lighting positions. This latent prior stabilizes highlights, and preserves background chromaticity, reducing color inconsistencies.

The material background and highlight features are fused via channel-wise attention and injected into a ControlNet [Zhang et al. 2023] to guide spatially consistent, lighting-stable, and color-faithful relighting. Combined with any multi-image SVBRDF framework, our method enables high-resolution SVBRDF recovery from a single image without baked-in highlights.

We validate our method on synthetic [Deschaintre et al. 2018; Vecchio and Deschaintre 2024] and real-world [Guo et al. 2020] material datasets against a state-of-the-art neural relighting baseline [Bieron

et al. 2023]. Quantitative and qualitative results show that our framework significantly improves SVBRDF recovery under strong highlights and generalizes across diverse materials. Code and pretrained models are available at <https://github.com/xingyouxin/DGRSISE>.

Our contributions can be summarized as follows:

- A novel ControlNet architecture based on stable diffusion generates relighting images under varying lighting positions, and can be integrated with pre-trained multi-image SVBRDF estimators to recover high-fidelity SVBRDF maps.
- A shuffle-based background consistency module that learns stable background features and provides implicit highlight region marks to remove baked-in artifacts.
- A specular prior reuse strategy that extracts highlight features and injects highlight priors, preserving both consistent specular and background color.

## 2 Related Work

This section reviews relevant work on neural relighting and surface SVBRDFs estimation.

### 2.1 Neural Relighting

Neural relighting methods fall into two categories: direct image-to-image mapping, which maps inputs to relighting results, and decomposition-based methods, which reconstruct geometry, materials, and lighting to improve consistency and editability.

*Direct image-to-image methods.* Early relighting methods captured dense reflectance fields under controlled illumination [Debevec et al. 2000], requiring specialized hardware. Neural methods now dominate by mapping images to relighting results. IC-Light [Zhang et al. 2025] fine-tunes a full-parameter conditional diffusion model with high training cost. DiLightNet [Zeng et al. 2024b] leverages ControlNet for efficient fine-tuning and radiance-guided illumination. Neural Gaffer [Jin et al. 2024] builds voxel-based scene representations for improved 3D consistency. In the field of surface materials, Bieron et al. [2023] proposed the first single-image neural relighting method, using residual and specular-aware modules without geometry or BRDF priors. Our method also targets material relighting, distinguished by introducing a feature extractor that disentangles background and highlights for robust conditioning.

*Decomposition-based methods.* Zeng et al. [2023] extend Neural Radiance Fields (NeRF) [Mildenhall et al. 2021] with shadow and highlight hints for realistic multi-view relighting. For single images, Zhu et al. [2022] use differentiable path tracing to jointly recover geometry, materials, and lighting, but at high computational cost. RGB $\leftrightarrow$ X [Zeng et al. 2024a] employs diffusion priors to separate reflectance and irradiance, enabling multimodal relighting and editing. DiffusionRenderer [Liang et al. 2025] integrates video diffusion with neural fields, exploiting generative priors for robust forward and inverse rendering under complex materials and lighting.

### 2.2 Surface SVBRDF Estimation

For brevity, we focus on images captured using lightweight devices (e.g., smartphones, tablets).

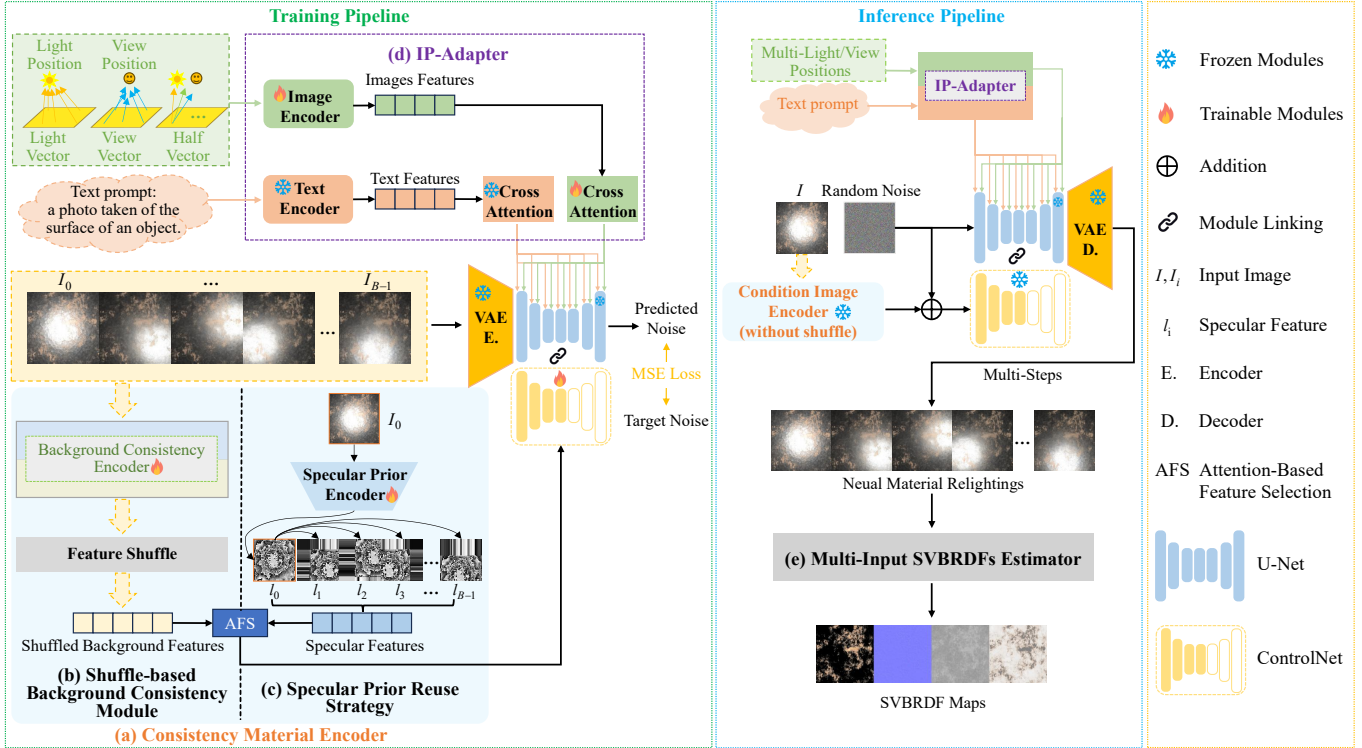


Fig. 2. Overview of our training and inference pipeline. **Training:** Given a batch of input images, we extract stable background features using the proposed (b) shuffle-based background consistency module (Section 4.1). In parallel, we use (c) a specular prior reuse strategy to translate highlight features  $l_0$  (encoded from the input  $I_0$ ) to novel lighting positions (Section 4.2). These two types of features are fused via a channel attention mechanism (AFS [Guo et al. 2021]) and injected into ControlNet [Zhang et al. 2023], providing structured and disentangled guidance for diffusion-based generation. Additionally, explicit information about the light vector, view vector, and their half vector is embedded into the cross-attention layers of the diffusion model using (d) an IP-Adapter mechanism [Ye et al. 2023], enabling precise and controllable illumination conditioning. **Inference:** Our method starts from a single input image  $I$ . Under different light and view positions, we use (a) a consistency material encoder without feature shuffle to extract features. These features guide the diffusion model to generate diverse neural material relighting results. By integrating with (e) a multi-input SVBRDF estimator [Luo et al. 2024b], it enables the reconstruction of high-quality SVBRDF maps.

*Single-input SVBRDF recovery.* Single-image SVBRDF estimation is convenient but ill-posed due to limited shading cues. Most methods use data-driven networks to infer reflectance from RGB inputs.

CNN-based methods leverage spatial priors for high-fidelity reconstruction, exploiting local self-similarity [Aittala et al. 2016; Henzler et al. 2021] and data augmentation [Li et al. 2017; Ye et al. 2018]. Highlight-aware modules [Guo et al. 2021], meta-learning [Zhou and Kalantari 2022], and gradient-based reflectance prediction [Luo et al. 2024a] further improve robustness. While preserving fine details, these methods are less effective for diverse material generation.

GAN-based methods employ adversarial learning to enhance generative diversity and conditional editing [Vecchio et al. 2021; Wen et al. 2022; Zhao et al. 2020; Zhou et al. 2022]. They offer higher editing flexibility but suffer from unstable training.

Diffusion-based methods offer strong generative priors and support multimodal guidance for diverse, controllable material synthesis. Vecchio et al. [2024] integrate CLIP-style conditioning with ControlNet to preserve spatial details, while Sartor and Peers [2023]

(MatFusion) introduce a flexible fine-tuning strategy updating only input layers. Despite their strong generative power, diffusion methods still struggle with strong highlights. Our consistency material encoder improves robustness under such conditions.

*Multi-input SVBRDF recovery.* Multi-view and multi-light inputs improve robustness to strong highlights. Early methods handle varying inputs via specialized networks [Deschaintre et al. 2019]. Others use cascaded networks for direct reflectance estimation [Kim et al. 2017]. Some exploit geometric priors, such as surface normals, to improve accuracy [Boss et al. 2020]. MaterialGAN [Guo et al. 2020] estimates SVBRDF from sparse images without extra priors by optimizing in a latent reflectance space. Recently, Luo et al. [2024b] use a graph convolutional network to capture cross-view correlations. They refine results with optimization-based fine-tuning. We adopt their pre-trained model as our SVBRDF estimator.

### 3 Neural Material Relighting Framework

Under strong lighting, single-image SVBRDF reconstruction often fails in highlight regions due to insufficient pixel information. These

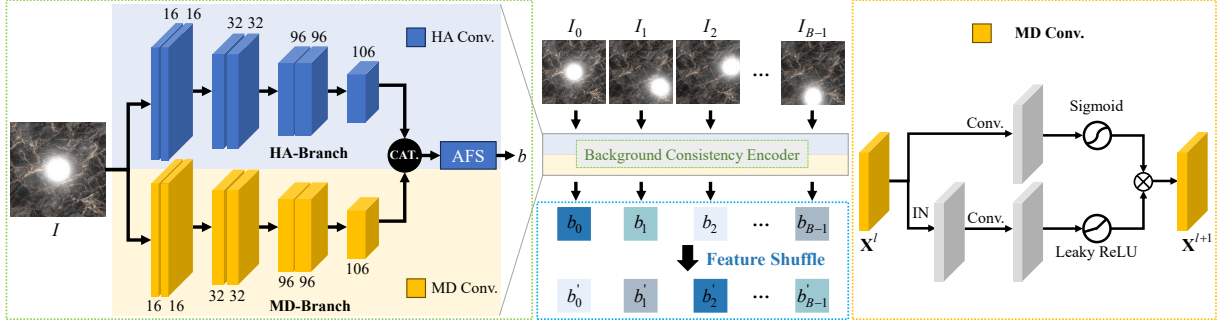


Fig. 3. Architecture of our proposed shuffle-based background consistency module. This figure illustrates the full pipeline for extracting stable background features (left), which consists of the HA-Branch and MD-Branch. It also depicts the intra-batch shuffling process used during training only to improve learning capacity and enforce consistency across features extracted from different lighting conditions (middle). The right part shows the details of the MD convolution. Here,  $\otimes$  denotes element-wise multiplication. IN indicates instance normalization.  $X^l, X^{l+1}$  represent the input and output of the  $l$ -th layer, respectively.

bake-in artifacts are reduced when images under varying highlights are available. To address this, we propose a neural material relighting framework. It synthesizes high-quality relighting images at diverse lighting and views from a single input, which can then be used for SVBRDF recovery.

Our framework is based on a ControlNet diffusion model with multi-modal conditioning. It has two key components. First, a consistency material encoder (Fig. 2 (a)) extracts disentangled representations of lighting and background to guide diffusion. Second, an IP-Adapter [Ye et al. 2023] (Fig. 2 (d)) encodes text and view/lighting conditions and injects them via cross-attention.

*Inference pipeline.* At inference, the framework takes three inputs: a single captured material image, a text prompt, and novel view/lighting positions. The consistency material encoder extracts stable background features and predicts a specular prior for the target light (Fig. 2, middle). These features are combined with a noise latent and passed to ControlNet. Text and view/lighting information are encoded by pretrained encoders and injected via cross-attention. Guided by these conditions, the diffusion model iteratively denoises to produce relighting results. Finally, the input image and seven relighting outputs are fed into Luo et al. [2024b]’s multi-image SVBRDF estimator to recover SVBRDF maps.

*Training pipeline.* The original ControlNet encoder is designed for low-level conditions such as normals, depth, or edges. It struggles with semantically richer inputs, like material images with strong specular highlights. To address this, we decouple background structure and highlight information, then recombine them in a task-aware way. During training, each batch contains renderings from the same SVBRDF maps under varying views and lighting (Fig. 2, left). The consistency material encoder has two parallel branches: the shuffle-based background consistency module (Section 4.1) and the specular prior reuse strategy (Section 4.2).

The capture setup of the material image follows the configuration in Aittala et al. [2015]. Light and view directions are parameterized by their position vectors and the half-vector. During training, light positions are randomly sampled within  $x, y \in [-4, 4]$  and  $z \in [4, 8]$  above the material plane. The view position is constrained above the

material center with random height  $z \in [4, 8]$ . An image encoder, trained jointly with the main network, extracts their features. For SVBRDF decomposition, we follow Luo et al. [2024b], using the Cook–Torrance microfacet BRDF [Cook and Torrance 1982] with the GGX normal distribution [Walter et al. 2007]. The SVBRDF is represented by four maps: diffuse albedo, surface normal, roughness, and specular albedo.

## 4 Consistency Material Encoder

Generating high-quality relighting under varying highlights requires accurate conditioning for consistent backgrounds and stable highlights. We address this with a material-specific feature extractor, the consistency material encoder. Inspired by rendering shading models, we decompose diffuse (background) and specular (highlight) components using a dual-branch network. The two streams disentangle features, which are fused via channel-wise attention [Guo et al. 2021] into a unified representation. This fused feature provides reliable priors for the diffusion model, producing stable and coherent relighting.

### 4.1 Shuffle-Based Background Consistency Module

Strong specular highlights in the input material image obscure underlying textures, making reliable background extraction difficult. To address this, we design a module for learning stable material backgrounds and implicit highlight marks under varying lighting, along with a shuffle-based training strategy (Fig. 3, middle). During training, a batch of images  $\{I_i \in \mathbb{R}^{B \times C \times H \times W} \mid i \in [0, B)\}$  of the same material under different lights is used. The module learns to extract consistent background features, where  $B$  is the batch size,  $C$  the number of channels, and  $H, W$  the spatial resolution.

Our architecture adapts the highlight-aware (HA) convolution and HA-Branch (Fig. 3, left) proposed by Guo et al. [2021], and introduces a material decomposition (MD) convolution (Fig. 3, right) and MD-Branch (Fig. 3, left). The MD convolution generates attention maps via a sigmoid activation,  $M = \sigma(\text{Conv}(F))$ , to modulate features and emphasize highlight-related activations:  $F_{\text{out}} = F \cdot M$ . The HA convolution has the same structure but adds an Inception component with two tracks: one  $3 \times 3$  convolution, the other two;

channels are halved and concatenated [Guo et al. 2021]. The Inception preserves global information but can reintroduce highlights. Without it, the MD-Branch focuses on stable background features. With it, the HA-Branch tends to predict specular highlight regions.

By encoding features through both HA and MD branches, the network separates highlight cues from diffuse structure. The HA-Branch localizes specular highlights, while the MD-Branch preserves geometric and textural consistency across lighting. The branches produce complementary features, which are concatenated and fused via the attention-based feature selection (AFS) [Guo et al. 2021]. The visualizations in Fig. 4 show the structure and texture of the extracted feature (b, e) and the specular highlight regions (c, f).

*Feature shuffle.* A set of material images from a fixed viewpoint under varying lighting share the same meso-scale geometry and structure but show different shading. To exploit this, we adopt a shuffle strategy that lets the network perceive shared appearance features during training, improving learning and generalization. The feature shuffle is applied only during training.

As shown in Fig. 3, a batch of training images  $\{I_i\}$  is first processed by the background consistency encoder (BCE) to extract background features  $\{b_i\}$ , where  $b_i = \text{BCE}(I_i)$ . The features are then shuffled by the feature shuffle operation (FS) to produce  $\{b'_i\}$ , where  $b'_i = \text{FS}(b_i)$ . This breaks the one-to-one correspondence between images and features, encouraging the network to learn stable background structures from a broader context during training.

Due to the non-generative nature of BCE, large specular highlights can partially obscure local features. To address this, we introduce auxiliary highlight marks to implicitly guide the diffusion model in generating coherent content within these regions (Fig. 4 (c, f)). This helps the model recover missing background and mitigate highlight bake-in artifacts. Validation results are in Section 6.3.

## 4.2 Specular Prior Reuse Strategy

We propose a specular prior reuse strategy to address limitations of using the shuffle-based background consistency module alone. Without it, the diffusion model’s latent representation lacks explicit specular highlight information and comprehensive background color priors. As a result, the model relies on learned statistics, causing inconsistent highlights and color shifts that deviate from the input.

To mitigate these issues, we reuse features extracted by a specular prior encoder from the input image, denoted as  $l_0$ . This encoder uses only the MD-Branch. The process is illustrated in Fig. 2 (c). During training,  $l_0$  is translated based on the given light positions, producing a set of features  $l_i$  for  $i \in [1, B]$ . Missing regions are filled using boundary values.

We employ the instance normalization (IN) branch in the MD convolution (Fig. 3, right) to enable the specular prior encoder to learn structural features more effectively. In parallel, the sigmoid branch predicts the probability of specular highlight regions. The element-wise product of these two branches yields features that preserve background information while capturing the highlight contours. Although feature translation may misalign the color features with the input at the pixel level, the diffusion model can still utilize these priors to reconstruct approximate color distributions. The effectiveness of this strategy is shown in Section 6.3.

## 5 Implementation Details

*Dataset.* We train on the MatSynth dataset [Vecchio and Deschaintre 2024], containing 3,980 high-resolution ( $2,048 \times 2,048$ ) material samples across categories like wood, terracotta, stone, plastic, marble, and fabric. Before training, each material is randomly scaled and cropped to generate 102,600 unique  $512 \times 512$  SVBRDF variants. During training, the maps are rendered with the Cook-Torrance model under the view and lighting setups in Section 3.

*Training and inference.* We finetune the pre-trained Stable Diffusion 2.1 [Rombach et al. 2022] with ControlNet using Diffusers [von Platen et al. 2022] and the Adam optimizer on a single Nvidia H100 GPU. The learning rate is  $5e-6$  with batch size 8. Training runs on  $512 \times 512$  resolution for 850,000 batches with mean square error (MSE) loss, taking about 5 days. At runtime, inference takes 1.6 seconds per image with a 20-step DDPM scheduler. Classifier-free guidance (CFG) [Ho and Salimans 2022] scale is 1, and other hyperparameters use Diffusers defaults.

For fairness, we retrained Bieron et al. [2023] using the same dataset, resolution, and view/lighting sampling as our method.

## 6 Results

In this section, we compare the quality of neural material relighting produced by our method against prior methods and demonstrate its improvements in single-image SVBRDF reconstruction.

### 6.1 Comparison on Neural Material Relighting

We compare our method with Bieron et al. [2023], the current state-of-the-art in neural material relighting. Quantitative results are summarized in Table 1. For each material in INRIA [Deschaintre et al. 2018] (40), MatSynth [Vecchio and Deschaintre 2024] (89), and real-world [Guo et al. 2020] (38) test sets, we generate seven relighting results under varying lights. Reconstruction errors are computed using MSE, root mean square error (RMSE), peak signal-to-noise ratio (PSNR), and learned perceptual image patch similarity (LPIPS). Averaged across all cases, our method outperforms Bieron et al. [2023] across INRIA and Real-world datasets, showing higher numerical accuracy and quality. Bieron et al. [2023] slightly outperforms our

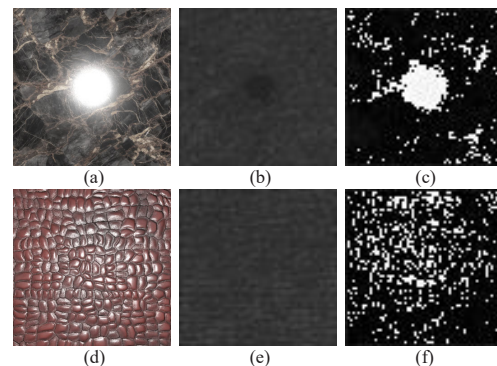


Fig. 4. Features extracted by the ST-Branch and MD-Branch from the input images. (a, d) show the input images; (b, e), and (c, f) show the channel-wise averaged visualizations of background structures and specular marks.

method on LPIPS for the MatSynth dataset, which contains more high-frequency details. Our method, limited by partial fine-tuning of the stable diffusion model, lacks such high-frequency details.

Table 1. Comparison of neural material relighting between our method and Bieron et al. [2023], with the best results per dataset highlighted in bold.

Dataset	Method	MSE↓	RMSE↓	PSNR↑	LPIPS↓
INRIA	Bieron	0.011	0.099	20.4	0.182
	Ours	<b>0.006</b>	<b>0.070</b>	<b>22.9</b>	<b>0.147</b>
MatSynth	Bieron	0.012	0.103	19.9	<b>0.213</b>
	Ours	<b>0.009</b>	<b>0.089</b>	<b>20.9</b>	0.244
Real-world	Bieron	0.018	0.126	18.6	0.210
	Ours	<b>0.014</b>	<b>0.112</b>	<b>19.2</b>	<b>0.166</b>

Fig. 6 provides visual comparisons under novel lighting. The first four materials are real-world [Guo et al. 2020], and the last two are synthetic from INRIA and MatSynth. Bieron et al. [2023] shows noticeable bake-in artifacts in highlight regions due to overfitting to the input lighting, especially under strong light. In contrast, our method is robust to highlight artifacts and produces stable, plausible specular responses, yielding relighting closer to the ground truth.

## 6.2 Improvements in Single-Image SVBRDF Recovery

We adopt the pretrained network of Luo et al. [2024b] as the multi-image SVBRDF estimator. A single input image and seven synthetic relighting images are fed into this estimator to recover SVBRDF maps. These maps are rendered under 128 view/lighting conditions for synthetic datasets and 7 for real-world data. Using this pipeline, we compare SVBRDF reconstruction quality against Bieron et al. [2023] and MatFusion [Sartor and Peers 2023] on both synthetic and real images. For baselines, we also evaluate Luo et al. [2024b] with (i) 1 ground-truth (GT) rendering input (baseline) and (ii) 8 GT rendering inputs (reference).

Table 2 presents the quantitative comparison. Our method consistently outperforms Bieron et al. [2023] and MatFusion [Sartor and Peers 2023] across most metrics except the LPIPS of the MatSynth dataset. And the reconstruction quality of our results lies between those achieved by the 1 GT and 8 GT input settings. Fig. 9 shows qualitative results: the first four materials are from real-world datasets, while the last two are drawn from the INRIA and MatSynth datasets. For specular materials, our recovered SVBRDFs are cleaner than both the 1 GT baseline and Bieron et al. [2023], and the rendered results more closely approximate those of the 8 GT reference and the synthetic GT. This demonstrates that our neural material relighting strategy can effectively enhance single-image SVBRDF reconstruction by enabling multi-image estimators.

Fig. 7 shows a qualitative comparison between our method, MatFusion [Sartor and Peers 2023], and LGD [Luo et al. 2024a]. Our method is robust to both spatially broad and strong specular highlights, recovering SVBRDFs that are free from highlight pollution. In contrast, due to intense lighting effects, MatFusion and LGD often produce baked-in highlights at the center of the estimated maps, resulting in visible artifacts in the relighting renderings. This

Table 2. Reconstruction and rendering error comparison between our method and others. SVBRDF maps are evaluated using MSE, and renderings are assessed with PSNR and LPIPS. D, N, R, S, and Ren. denote diffuse, normal, roughness, specular, and re-rendered results, respectively. For each dataset, the best results are in bold and the second-best are underlined. N/A indicates results not applicable, as ground-truth SVBRDF maps are unavailable for the real-world dataset.

Dataset	Method	MSE↓				PSNR↑	LPIPS↓
		D	N	R	S	Ren.	Ren.
INRIA	Bieron	0.011	0.005	0.050	0.038	20.0	0.206
	MatFusion	0.007	0.004	0.068	0.015	19.1	0.253
	Ours	<u>0.006</u>	<u>0.003</u>	<u>0.018</u>	<u>0.024</u>	<u>22.9</u>	<u>0.187</u>
	1 GT	0.008	0.006	0.020	<u>0.014</u>	21.4	0.206
	8 GT	<b>0.003</b>	<b>0.001</b>	<b>0.006</b>	<b>0.009</b>	<b>27.8</b>	<b>0.085</b>
MatSynth	Bieron	0.019	0.009	0.043	0.074	20.1	0.249
	MatFusion	0.018	0.009	0.080	0.078	18.7	<u>0.239</u>
	Ours	<u>0.017</u>	<u>0.008</u>	<u>0.018</u>	<u>0.068</u>	<u>20.5</u>	0.312
	1 GT	0.018	0.012	0.030	0.079	20.4	0.270
	8 GT	<b>0.012</b>	<b>0.003</b>	<b>0.010</b>	<b>0.038</b>	<b>25.1</b>	<b>0.107</b>
Real-world	Bieron	N/A	N/A	N/A	N/A	18.4	0.226
	MatFusion	N/A	N/A	N/A	N/A	15.8	0.282
	Ours	N/A	N/A	N/A	N/A	<u>19.5</u>	<u>0.199</u>
	1 GT	N/A	N/A	N/A	N/A	18.4	0.233
	8 GT	N/A	N/A	N/A	N/A	<b>22.1</b>	<b>0.164</b>

demonstrates that the ability of our method to handle specular highlights arises from the specialized network design rather than from the diffusion model.

## 6.3 Ablation Study

We use the original ControlNet encoder as the baseline (Ori. CtrlNet) and compare it against our full method and several ablated variants. For evaluation, each method generates seven neural material relighting results per sample on the MatSynth test set, which contains 89 samples. We compute MSE, RMSE, PSNR, structural similarity index measure (SSIM), and LPIPS between the relighting results and their ground truth. Detailed numerical results are reported in Table 3. To better illustrate the perceptual differences between the relighting results and the ground truth, we further employ the FLIP metric [Andersson et al. 2020] for visual error quantification. A full visual comparison is presented in Fig. 8.

*The effect of the shuffle-based BCE.* In Fig. 8, we compare relighting results with SBCE (ours) and without it (w/o SBCE), where SBCE consists of both shuffling and the BCE module. The inclusion of this module leads to significantly improved preservation of the example material’s inherent grid-like texture patterns and base color characteristics. This improvement arises from the module’s design: by leveraging a shuffle strategy, the MD and HA branches jointly capture the global structural context of the material background, which serves as a stable conditioning signal for the diffusion-based relighting process.

*The effect of the HA and MD convolution.* To evaluate the effect of HA and MD convolutions, we replace all convolution layers with standard convolutions and denote this variant as w/o HA&MD. As shown in Fig. 8, using only standard convolutions leads to a significant degradation in the material’s base color fidelity. This is

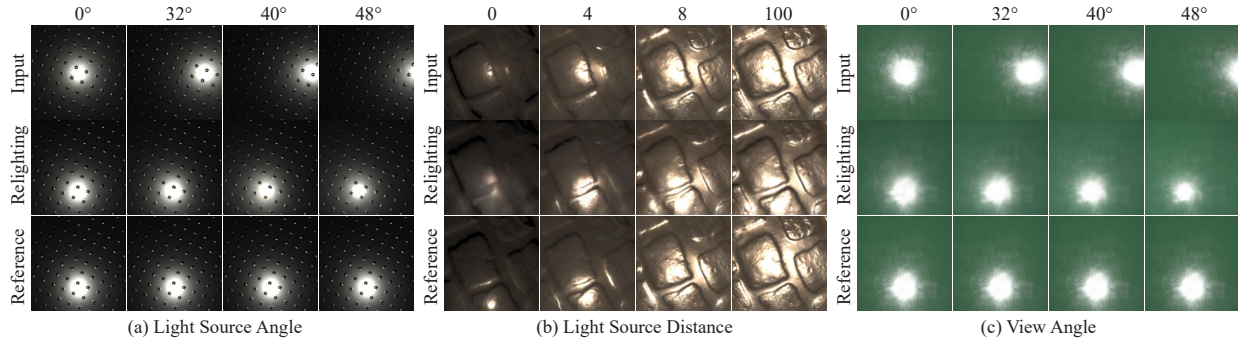


Fig. 5. Our neural material relighting is robust to input parameters. (a) Varying light source angles. (b) Varying light source distances. (c) Varying view angles.

primarily because the absence of Inception and IN modules in the HA and MD slows color learning convergence. Second, the use of a shuffling strategy disrupts color features by swapping their order, causing the network to focus on easier-to-learn global structures and neglect color information. In addition, relying solely on standard convolutions results in baked-in artifacts at the center of the relighting output.

*The effect of the specular prior reuse strategy.* We assess the effectiveness of the specular prior reuse strategy by comparing relighting results with specular prior (ours) and without it (w/o S. Prior). As shown in Table 3 and Fig. 8, incorporating the specular prior leads to more faithful highlight placement and shape, with background color more closely aligned with the ground truth. These improvements indicate that the specular prior encoder successfully captures the global background features and implicit highlight contours, providing enhanced conditioning for relighting.

Table 3. Numerical evaluation of our full method and ablated variants for neural material relighting. The best results are highlighted in bold.

Method	MSE↓	RMSE↓	PSNR↑	SSIM↑	LPIPS↓
Ori. CtrlNet	0.030	0.159	15.9	0.388	0.327
w/o SBCE	0.023	0.141	17.0	0.227	0.525
w/o HA&MD	0.018	0.120	18.6	0.421	0.299
w/o S. Prior	0.010	0.090	20.7	0.434	0.288
Ours	<b>0.009</b>	<b>0.089</b>	<b>20.9</b>	<b>0.442</b>	<b>0.244</b>

*Relighting images and the convergence accuracy of SVBRDF.* We compare SVBRDF re-rendering results using 1, 4, and 8 relighting inputs, as shown in Table 4. Increasing the number of relighting images leads to improved SVBRDF convergence accuracy.

*Relighting robustness to input parameters.* When capturing material input images in real-world conditions, physical constraints inevitably introduce a positional discrepancy between the actual view/lighting configuration and the nominal setup provided to the network. To evaluate the sensitivity to such deviations, we conduct robustness tests. As shown in Fig. 5, our relighting is robust to lighting angles ( $0^\circ$ – $32^\circ$ ), light distances (4–8 units), and camera angles ( $0^\circ$ – $32^\circ$ ).

Table 4. Effect of the number of relighting inputs on re-rendering results, evaluated on the INRIA dataset. The best results are highlighted in bold.

Method	MSE↓	PSNR↑	LPIPS↓
1-input	0.0075	21.36	0.2059
4-input	0.0057	22.46	0.1874
8-input	<b>0.0052</b>	<b>22.88</b>	<b>0.1868</b>

#### 6.4 Limitations and Discussion

We have identified several limitations. First, the large scale of the Stable Diffusion 2.1 and ControlNet models makes training resource-intensive, requiring over 30 GB of VRAM and significant training time. For inference, efficiency is heavily bottlenecked by high resolutions and the denoising steps. Second, due to the regularizing effect of the KL divergence in the VAE framework [Wang et al. 2023] and partial parameter fine-tuning, the diffusion model tends to attenuate high-frequency details during generation. This leads to visibly reduced contrast and a lack of fine-grained textural fidelity, particularly for diffuse materials. Finally, our relighting is confined to a limited set of view/lighting conditions and does not support dynamic changes in illumination intensity. While this design choice limits expressiveness, it greatly simplifies training and provides enough inductive priors that facilitate effective SVBRDFs disentanglement.

#### 7 Conclusion

In this paper, we present a novel framework for high-fidelity neural material relighting. The core of our approach consists of a shuffle-based BCE and a specular prior reuse strategy. We introduce MD-Branch that captures background structural feature, and HA-Branch that predicts specular region marks, both of which provide reliable conditioning signals to the ControlNet for stable guidance. The shuffle strategy enhances the model’s awareness of global material geometry and structural context. Additionally, our specular prior reuse strategy injects learned highlight contours into new spatial locations and preserves material color priors, significantly improving relighting quality. Benefiting from these components, our framework demonstrates strong robustness to intense specularities across both real and synthetic datasets.

Future work may explore full fine-tuning or adopting stronger base models to improve resolution, detail fidelity. Another direction lies in extending the material model to better support complex recovery cases such as car paint and beetle shells, which exhibit anisotropic or multilayered reflectance.

## Acknowledgments

We thank the reviewers for the valuable comments. This work has been partially supported by the National Key R&D Program of China under grant No. 2022YFB3303203 and National Natural Science Foundation of China under grant No. 62272275, 62172220 and 62572230.

## References

- Miika Aittala, Timo Aila, and Jaakko Lehtinen. 2016. Reflectance modeling by neural texture synthesis. *ACM Transactions on Graphics (ToG)* 35, 4 (2016), 1–13.
- Miika Aittala, Tim Weyrich, Jaakko Lehtinen, et al. 2015. Two-shot SVBRDF capture for stationary materials. *ACM Trans. Graph.* 34, 4 (2015), 110–1.
- Pontus Andersson, Jim Nilsson, Tomas Akenine-Möller, Magnus Oskarsson, Kalle Åström, and Mark D Fairchild. 2020. FLIP: A Difference Evaluator for Alternating Images. *Proc. ACM Comput. Graph. Interact. Tech.* 3, 2 (2020), 15–1.
- James Bieron, Xin Tong, and Pieter Peers. 2023. Single Image Neural Material Relighting. In *ACM SIGGRAPH 2023 Conference Proceedings*. 1–11.
- Mark Boss, Varun Jampani, Kihwan Kim, Hendrik Lensch, and Jan Kautz. 2020. Two-shot spatially-varying brdf and shape estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3982–3991.
- Robert L Cook and Kenneth E. Torrance. 1982. A reflectance model for computer graphics. *ACM Transactions on Graphics (ToG)* 1, 1 (1982), 7–24.
- Paul Debevec, Tim Hawkins, Chris Tchou, Haarm-Pieter Duiker, Westley Sarokin, and Mark Sagar. 2000. Acquiring the reflectance field of a human face. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*. 145–156.
- Valentin Deschaintre, Miika Aittala, Fredo Durand, George Drettakis, and Adrien Bousseau. 2018. Single-image svbrdf capture with a rendering-aware deep network. *ACM Transactions on Graphics (ToG)* 37, 4 (2018), 1–15.
- Valentin Deschaintre, Miika Aittala, Frédo Durand, George Drettakis, and Adrien Bousseau. 2019. Flexible SVBRDF Capture with a Multi-Image Deep Network. In *Computer Graphics Forum*, Vol. 38.
- Duan Gao, Xiao Li, Yue Dong, Pieter Peers, Kun Xu, and Xin Tong. 2019. Deep inverse rendering for high-resolution SVBRDF estimation from an arbitrary number of images. *ACM Trans. Graph.* 38, 4 (2019), 134–1.
- Jie Guo, Shuichang Lai, Chengzhi Tao, Yuelong Cai, Lei Wang, Yanwen Guo, and Ling-Qi Yan. 2021. Highlight-aware two-stream network for single-image SVBRDF acquisition. *ACM Transactions on Graphics (TOG)* 40, 4 (2021), 1–14.
- Jie Guo, Shuichang Lai, Qinghao Tu, Chengzhi Tao, Changqing Zou, and Yanwen Guo. 2023. Ultra-high resolution svbrdf recovery from a single image. *ACM Transactions on Graphics* 42, 3 (2023), 1–14.
- Yu Guo, Cameron Smith, Miloš Hašan, Kalyan Sunkavalli, and Shuang Zhao. 2020. MaterialGAN: Reflectance capture using a generative SVBRDF model. *arXiv preprint arXiv:2010.00114* (2020).
- Philipp Henzler, Valentin Deschaintre, Niloy J Mitra, and Tobias Ritschel. 2021. Generative Modelling of BRDF Textures from Flash Images. *ACM Transactions on Graphics* 40, 6 (2021).
- Jonathan Ho and Tim Salimans. 2022. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598* (2022).
- Haian Jin, Yuan Li, Fujun Luan, Yuanbo Xiangli, Sai Bi, Kai Zhang, Zexiang Xu, Jin Sun, and Noah Snavely. 2024. Neural gaffer: Relighting any object via diffusion. *Advances in Neural Information Processing Systems* 37 (2024), 141129–141152.
- Kihwan Kim, Jinwei Gu, Stephen Tyree, Pavlo Molchanov, Matthias Nießner, and Jan Kautz. 2017. A lightweight approach for on-the-fly reflectance estimation. In *Proceedings of the IEEE International Conference on Computer Vision*. 20–28.
- Xiao Li, Yue Dong, Pieter Peers, and Xin Tong. 2017. Modeling surface appearance from a single photograph using self-augmented convolutional neural networks. *ACM Transactions on Graphics (ToG)* 36, 4 (2017), 1–11.
- Ruofan Liang, Zan Gojic, Huan Ling, Jacob Munkberg, Jon Hasselgren, Chih-Hao Lin, Jun Gao, Alexander Keller, Nandita Vijaykumar, Sanja Fidler, et al. 2025. Diffusion Renderer: Neural Inverse and Forward Rendering with Video Diffusion Models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 26069–26080.
- Di Luo, Hanxiao Sun, Lei Ma, Jian Yang, and Beibei Wang. 2024b. Correlation-aware Encoder-Decoder with Adapters for SVBRDF Acquisition. In *SIGGRAPH Asia 2024 Conference Papers*. 1–10.
- Xuejiao Luo, Leonardo Scandolo, Adrien Bousseau, and Elmar Eisemann. 2024a. Single-Image SVBRDF Estimation with Learned Gradient Descent. In *Computer graphics forum*, Vol. 43. Wiley Online Library, e15018.
- Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. 2021. Nerf: Representing scenes as neural radiance fields for view synthesis. *Commun. ACM* 65, 1 (2021), 99–106.
- Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 10684–10695.
- Sam Sartor and Pieter Peers. 2023. Matfusion: a generative diffusion model for svbrdf capture. In *SIGGRAPH Asia 2023 conference papers*. 1–10.
- Giuseppe Vecchio and Valentin Deschaintre. 2024. Matsynth: A modern pbr materials dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 22109–22118.
- Giuseppe Vecchio, Rosalie Martin, Arthur Roullier, Adrien Kaiser, Romain Rouffet, Valentin Deschaintre, and Tamy Boubekeur. 2024. Controlmat: a controlled generative approach to material capture. *ACM Transactions on Graphics* 43, 5 (2024), 1–17.
- Giuseppe Vecchio, Simone Palazzo, and Concetto Spampinato. 2021. Surfacenet: Adversarial svbrdf estimation from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 12840–12848.
- Patrick von Platen, Suraj Patil, Anton Lozhkov, Pedro Cuenca, Nathan Lambert, Kashif Rasul, Mishig Davaadorj, Dhruv Nair, Sayak Paul, William Berman, Yiyi Xu, Steven Liu, and Thomas Wolf. 2022. Diffusers: State-of-the-art diffusion models. <https://github.com/huggingface/diffusers>.
- Bruce Walter, Stephen R Marschner, Hongsong Li, and Kenneth E Torrance. 2007. Microfacet Models for Refraction through Rough Surfaces. *Rendering techniques 2007* (2007), 18th.
- Yikai Wang, Chenjie Cao, and Ke Fan Xiangyang Xue Yanwei Fu. 2023. Towards Context-Stable and Visual-Consistent Image Inpainting. *arXiv preprint arXiv:2312.04831* (2023).
- Tao Wen, Beibei Wang, Lei Zhang, Jie Guo, and Nicolas Holzschuch. 2022. SVBRDF Recovery from a Single Image with Highlights Using a Pre-trained Generative Adversarial Network. In *Computer Graphics Forum*, Vol. 41. Wiley Online Library, 110–123.
- Hu Ye, Jun Zhang, Sibao Liu, Xiao Han, and Wei Yang. 2023. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721* (2023).
- Wenjie Ye, Xiao Li, Yue Dong, Pieter Peers, and Xin Tong. 2018. Single image surface appearance modeling with self-augmented cnns and inexact supervision. In *Computer Graphics Forum*, Vol. 37. Wiley Online Library, 201–211.
- Chong Zeng, Guojun Chen, Yue Dong, Pieter Peers, Hongzhi Wu, and Xin Tong. 2023. Relighting neural radiance fields with shadow and highlight hints. In *ACM SIGGRAPH 2023 Conference Proceedings*. 1–11.
- Chong Zeng, Yue Dong, Pieter Peers, Youkang Kong, Hongzhi Wu, and Xin Tong. 2024b. DiLightNet: Fine-grained lighting control for diffusion-based image generation. In *ACM SIGGRAPH 2024 Conference Papers*. 1–12.
- Zheng Zeng, Valentin Deschaintre, Iliyan Georgiev, Yannick Hold-Geoffroy, Yiwei Hu, Fujun Luan, Ling-Qi Yan, and Miloš Hašan. 2024a. RGB $\leftrightarrow$ X: Image decomposition and synthesis using material- and lighting-aware diffusion models. In *ACM SIGGRAPH 2024 Conference Papers* (Denver, CO, USA) (SIGGRAPH '24). Association for Computing Machinery, New York, NY, USA, Article 75, 11 pages. doi:10.1145/3641519.3657445
- Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. 2023. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*. 3836–3847.
- Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. 2025. Scaling in-the-wild training for diffusion-based illumination harmonization and editing by imposing consistent light transport. In *The Thirteenth International Conference on Learning Representations*.
- Yezi Zhao, Beibei Wang, Yanning Xu, Zheng Zeng, Lu Wang, and Nicolas Holzschuch. 2020. Joint SVBRDF Recovery and Synthesis From a Single Image using an Unsupervised Generative Adversarial Network. In *EGSR (DL)*. 53–66.
- Xilong Zhou, Milos Hasan, Valentin Deschaintre, Paul Guerrero, Kalyan Sunkavalli, and Nima Khademi Kalantari. 2022. Tilegen: Tileable, controllable material generation and capture. In *SIGGRAPH Asia 2022 conference papers*. 1–9.
- Xilong Zhou and Nima Khademi Kalantari. 2022. Look-ahead training with learned reflectance loss for single-image svbrdf estimation. *ACM Transactions on Graphics (TOG)* 41, 6 (2022), 1–12.
- Jingsen Zhu, Fujun Luan, Yuchi Huo, Zihao Lin, Zhihua Zhong, Dianbing Xi, Rui Wang, Hujun Bao, Jiayang Zheng, and Rui Tang. 2022. Learning-based inverse rendering of complex indoor scenes with differentiable monte carlo raytracing. In *SIGGRAPH Asia 2022 Conference Papers*. 1–8.

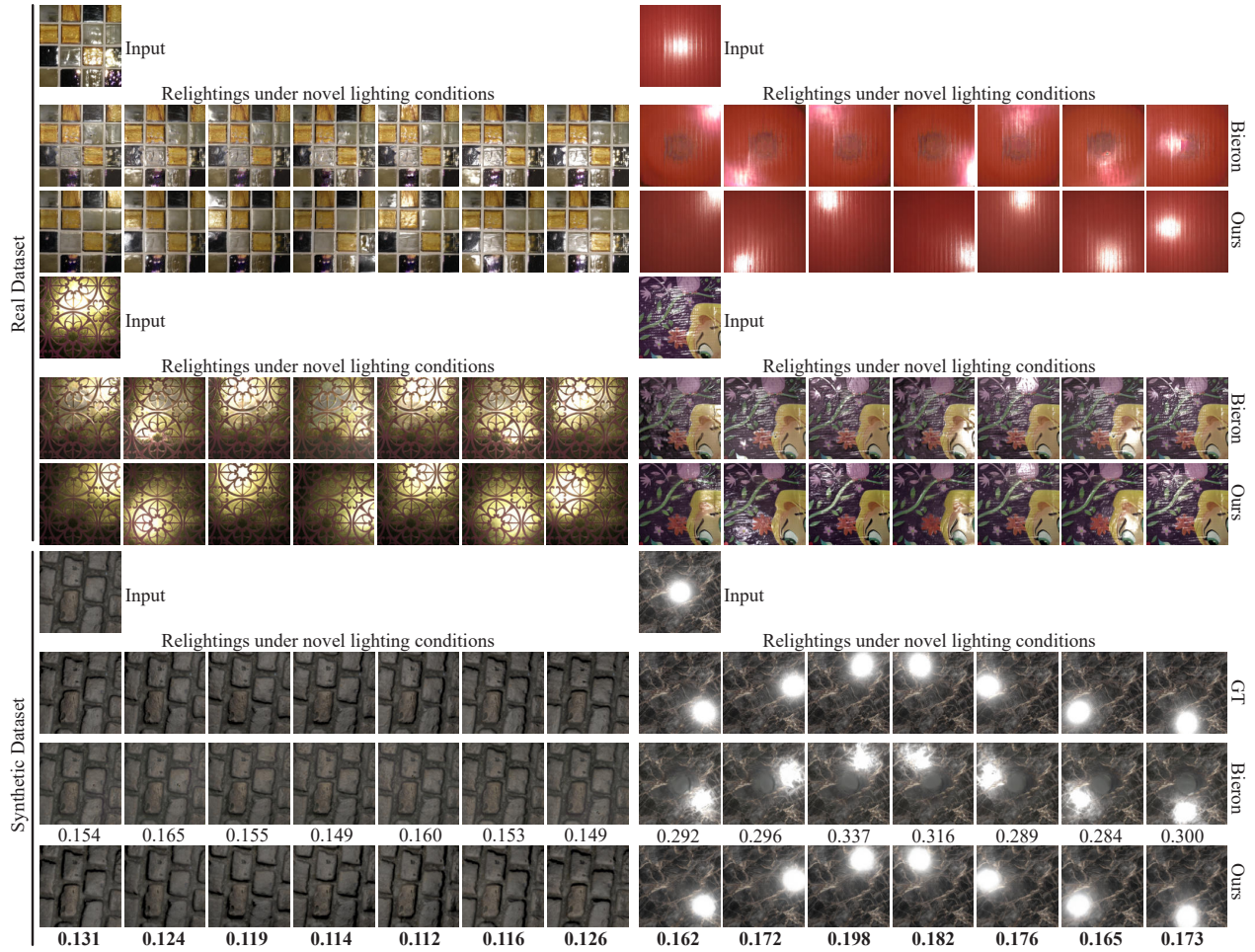


Fig. 6. Visual comparison of neural material relighting on real and synthetic datasets. We compare our method against [Bieron et al. \[2023\]](#) under the consistent view/lighting settings during training. For synthetic data, LIPS scores are reported below each image, with the lowest values highlighted in bold.

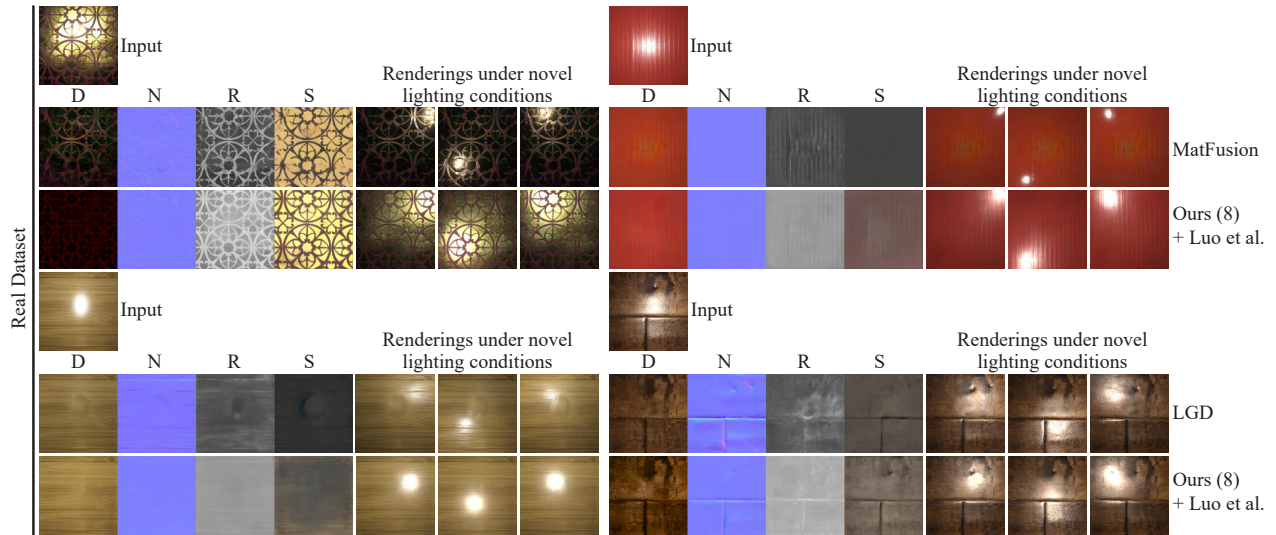


Fig. 7. Visual comparison of SVBRDFs and re-rendered results between our method, MatFusion ([Sartor and Peers \[2023\]](#)), and LGD ([Luo et al. \[2024a\]](#)) on real captured data. Our method effectively avoids specular bake-in artifacts in the SVBRDFs and produces clean renderings under novel lighting conditions.

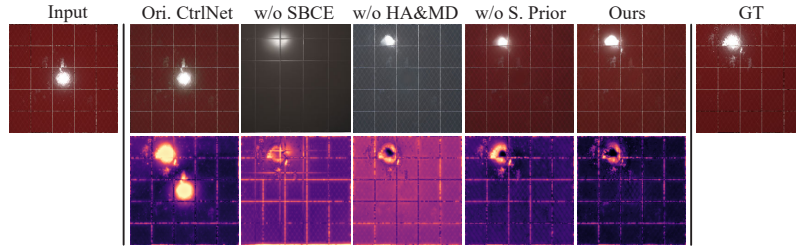


Fig. 8. Visual comparison of our full method and ablated variants for neural relighting. Compared to ours, w/o S. Prior exhibits slight color shifts.

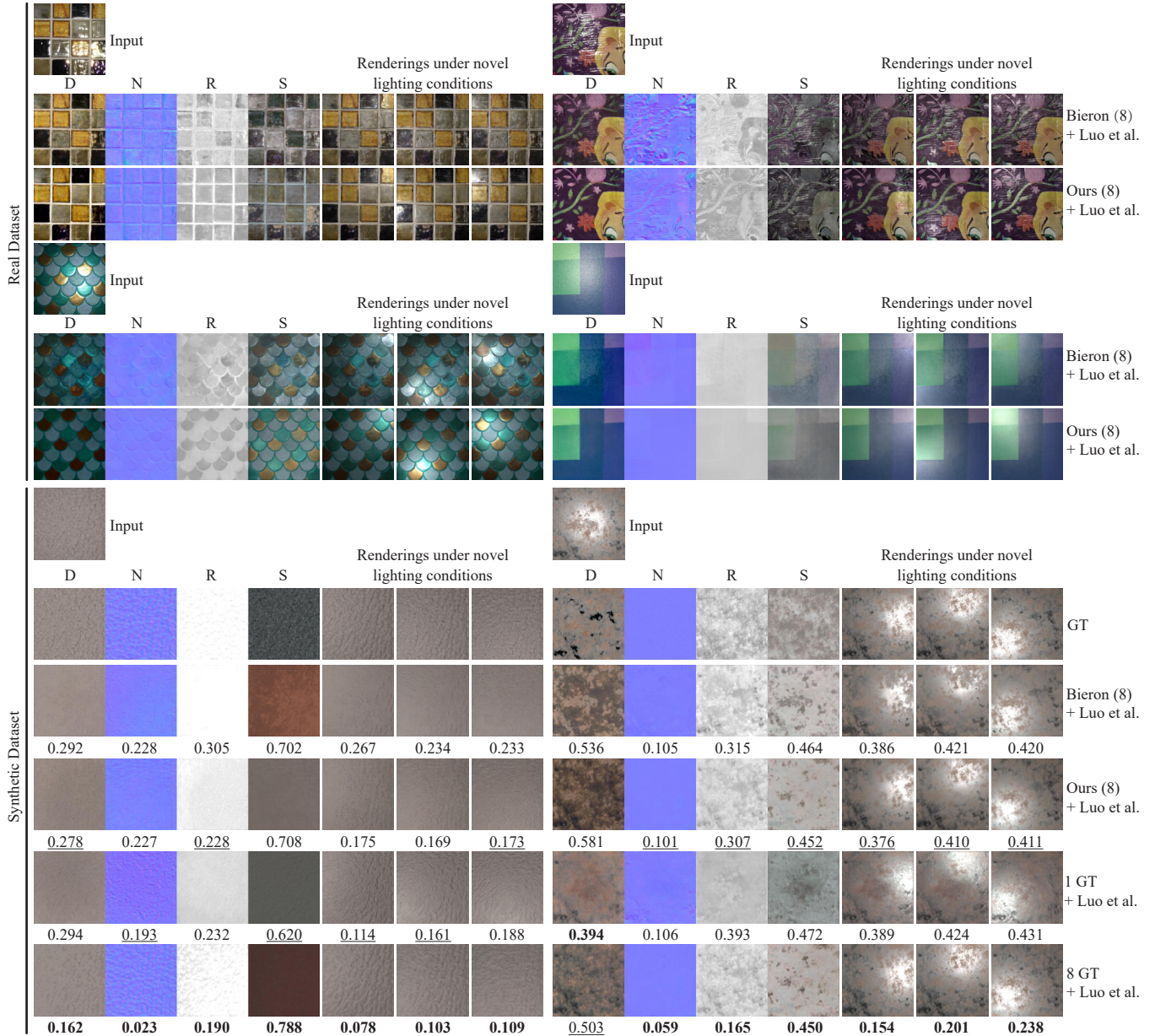


Fig. 9. Visual comparison of SVBRDFs and re-rendered results on real and synthetic datasets. We compare our method with Bieron et al. [2023] under the consistent view/lighting settings during training. For synthetic data, LPIPS scores are reported below each image, with the lowest and second lowest values highlighted in bold and underlined.