This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

MVInverse: Feed-forward Multiview Inverse Rendering in Seconds

Xiangzuo Wu[1] Chengwei Ren[2] Jun Zhou[1] Xiu Li[1†] Yuan Liu[2†]
[1]Tsinghua University [2]The Hong Kong University of Science and Technology

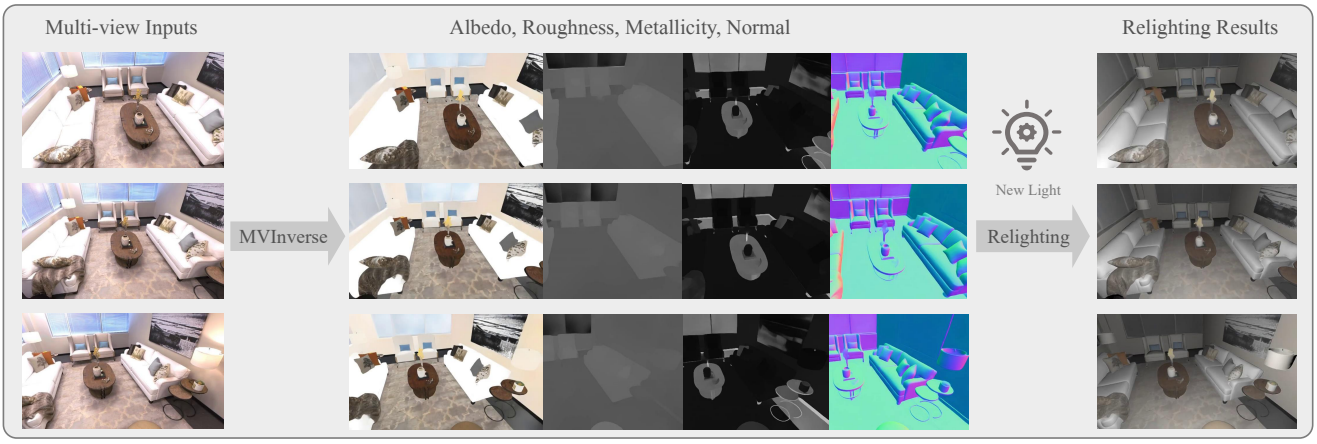

Figure 1. We present **MVInverse**, a feed-forward multi-view inverse rendering framework that jointly recovers consistent geometry and material properties from input images or videos. With the recovered properties, realistic relighting under novel illumination can be achieved within seconds.

Abstract

Multi-view inverse rendering aims to recover geometry, materials, and illumination consistently across multiple viewpoints. When applied to multi-view images, existing single-view approaches often ignore cross-view relationships, leading to inconsistent results. In contrast, multi-view optimization methods rely on slow differentiable rendering and per-scene refinement, making them computationally expensive and hard to scale. To address these limitations, we introduce a feed-forward multi-view inverse rendering framework that directly predicts spatially varying albedo, metallic, roughness, diffuse shading, and surface normals from sequences of RGB images. By alternating attention across views, our model captures both intra-view long-range lighting interactions and inter-view material consistency, enabling coherent scene-level reasoning within a single forward pass. Due to the scarcity of real-world training data, models trained on existing synthetic datasets often struggle to generalize to real-world scenes. To overcome this limitation, we propose a consistency-based fine-tuning strategy that leverages unlabeled real-world videos to enhance both multi-view coherence and robustness under in-the-wild conditions. Extensive experiments on benchmark datasets demonstrate that our method achieves state-of-the-art performance in terms of multi-view consistency, material and normal estimation quality, and generalization to real-world imagery. Code, models and more results are available at https://maddog241.github.io/mvinverse-page/

1. Introduction

Inverse rendering, a fundamental task in computer vision and graphics, aims to recover intrinsic scene properties—including albedo, roughness, and metallicity—from visual observations. This capability is pivotal across numerous practical domains: in film and animation production, it enables efficient digital asset creation by reconstructing physically accurate materials from real-world captures [98, 100, 119]; in augmented reality (AR) and virtual reality (VR), it facilitates seamless integration of virtual objects with real environments by matching their material properties to the scene [17, 74, 78]; and in robotics, it enhances scene understanding for tasks like object manipulation by inferring physical characteristics of surfaces [14, 101, 112]. Beyond these applications, inverse rendering

†Corresponding author

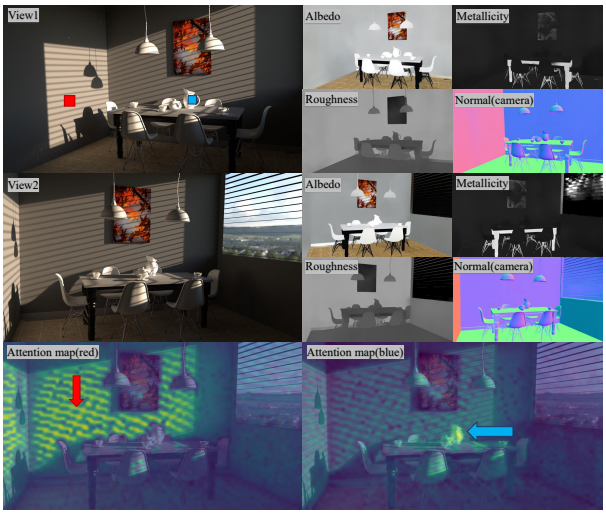

Figure 2. **Effect of Alternating Attention for Inverse Rendering**. Given the red and blue query patches in the first view, we visualize the corresponding attention heatmaps in the second view (bottom row) to illustrate the effectiveness of our model design. Red: the model captures long-range lighting interactions across spatially distant regions. Blue: the model maintains cross-view consistency by correctly associating corresponding surface regions under viewpoint changes.

serves as a cornerstone for advancing physically based rendering (PBR) pipelines, bridging the gap between image-based perception and 3D scene reconstruction.

Despite its significance, existing inverse rendering methods face critical limitations that hinder their practical deployment. First, the majority of approaches rely on iterative optimization frameworks [2, 9, 33, 42, 46, 62, 72, 87, 96, 109, 117], which often yield high-fidelity results but suffer from prohibitive computational costs, requiring minutes to hours of processing per scene, making them infeasible for real-time or large-scale applications. Second, while some single-view intrinsic decomposition methods [16, 53, 55–57, 122] achieve fast inference by leveraging feed-forward networks, their extension to multi-view settings is plagued by multi-view inconsistency: material predictions for the same 3D region across different viewpoints often diverge, leading to unrealistic artifacts in reconstructed scenes. Third, recent diffusion-based renderers have demonstrated promising results for image and video inverse rendering [48, 60, 111], but their inherent generative nature and iterative sampling process render them computationally expensive, inheriting the speed limitations of optimization-based methods. Thus, developing an inverse rendering solution that balances speed and accuracy remains an open and pressing challenge.

Our work is motivated by the remarkable recent progress in feed-forward 3D reconstruction, where vision transformer (ViT)-based models have revolutionized the field by replacing lengthy optimization pipelines with direct, end-to-end prediction [44, 67, 90, 93]. Multi-view inverse rendering shares a key similarity with these reconstruction methods: both traditionally depend on repeated rendering iterations to solve for intrinsic properties (materials for inverse rendering, geometry for reconstruction). This parallel insight suggests that integrating the feed-forward paradigm into inverse rendering is a natural and unexplored direction. By leveraging the efficiency of transformer-based architectures to directly map multi-view inputs to material properties, we can potentially overcome the speed bottlenecks of optimization and diffusion-based methods while addressing the consistency issues of monocular approaches.

Building on this motivation, we propose **MVInverse**, a novel feed-forward model that predicts albedo, roughness, and metallicity along with camera-space normals and diffuse shading maps for all input views in just a few seconds. Inspired by recent state-of-the-art feed-forward 3D reconstruction architectures [90, 93], we adopt a VGGT-style [90] transformer backbone with alternating global and frame-wise attention: the global attention explicitly associates corresponding 3D regions across different viewpoints, while the frame-wise attention captures intra-frame interactions in individual images. This alternating-attention design extracts efficient, viewpoint-aligned features that serve as a robust foundation for subsequent material and normal prediction. As illustrated in Figure 2, attention visualizations reveal that our model effectively captures both long-range lighting dependencies (red patch) and inter-frame consistency (blue patch), highlighting the dual capability of the alternating attention mechanism.

A further challenge in inverse rendering is the scarcity of high-quality real-world training data. Most existing methods are trained on synthetic datasets [25, 51, 52, 79, 92, 120, 122], leading to poor generalization to real-world scenarios, particularly the emergence of flickering artifacts, where material properties of the same 3D region fluctuate inconsistently across views or frames. To mitigate this, we adopt a two-stage training strategy: we first pre-train our model on large-scale synthetic datasets to learn general material priors, then introduce a self-supervised consistency constraint during fine-tuning on real-world data. This constraint enforces that material predictions for the same 3D point, as observed from different viewpoints, remain consistent, significantly improving the model's generalization to real environments and eliminating flickering artifacts.

We validate our approach across common benchmark datasets, demonstrating strong single-view inverse rendering quality, multi-view material consistency, accurate normal prediction, and stable temporal performance on videos. Moreover, by combining our method with existing feed-forward 3D reconstruction pipelines, we can obtain textured point clouds that enable various downstream appli-

cations such as video relighting and view-consistent material editing in only a few seconds. This efficient end-to-end pipeline greatly lowers the barrier for practical deployment, paving the way for highly scalable, data-driven material understanding.

2. Related Work

Single-view Inverse Rendering Inverse rendering aims to decompose a scene into geometry, material properties and illumination [6]. This is an inherently ill-posed task as multiple combinations of these properties can produce the same image. Early approaches relied on handcrafted priors and physical heuristics to constrain the problem [4, 6, 7, 31, 36, 50, 82, 83, 115, 118], which were often limited by the applicability of their underlying assumptions to diverse real-world scenes. The advent of deep learning has shifted the paradigm towards data-driven techniques. Modern methods often frame inverse rendering as an image-to-image translation problem, training dense prediction neural networks on large-scale synthetic datasets with ground truth decompositions [15, 16, 53, 55–57, 70, 75, 80, 81, 84, 94, 95, 110, 113, 116, 122, 123]. These learning-based approaches can jointly estimate spatially-varying bidirectional reflectance distribution functions (SVBRDFs), complex lighting, and surface normals. More recently, generative models, particularly diffusion models, have been employed to address the problem in a formulation of conditional image generation [28, 29, 48, 59–61, 76, 102, 111, 121]. These models leverage powerful priors learned from large-scale image distributions to synthesize physically plausible decompositions. However, a fundamental limitation of all single-view methods is their inability to enforce geometric and material consistency across different viewpoints of the same scene. In contrast, our proposed method leverages multiple views simultaneously to resolve this ambiguity and ensure a coherent scene decomposition.

Multi-view Inverse Rendering Multi-view inverse rendering leverages information from multiple observations of a scene to enable more accurate recovery of geometry, materials, and illumination. Early multi-view methods often relied on explicit geometry and heavy optimization to jointly refine these factors [2, 41, 72, 96, 109], achieving physically accurate results through iterative optimization. With the rise of neural scene representations, NeRF [71] and NeuS [91]-based approaches [1, 9–11, 13, 21, 46, 66, 87, 89, 103–105, 114, 117] have become a dominant paradigm. These methods model radiance and reflectance within implicit neural fields and typically optimize network parameters per scene to disentangle view-dependent appearance and incident illumination. More recently, 3D Gaussian Splatting [45] has been adapted for inverse rendering [8, 18, 24, 27, 33, 37, 38, 42, 49, 62, 85, 97, 107, 108, 124], offering real-time differentiable rendering and efficient geometry representation with some feed-forward 3D geometry reconstruction [30, 34, 68, 90, 93]. In addition, several hybrid methods [20, 63, 65, 69] adopt an estimation-then-optimization strategy, where learning-based predictions are refined through per-scene optimization. Despite recent progress, most existing approaches still rely on scene-specific optimization or fine-tuning to achieve material and lighting consistency. Although learning-based systems such as MAIR and MAIR++ [22, 23] consider multi-view settings without optimization, their applicability remains limited to relatively constrained viewpoints. In contrast, our method performs multi-view inverse rendering in a single feed-forward pass, producing coherent scene decomposition even under large camera motions.

3. Method

In this section, we introduce MVInverse, a feed-forward inverse rendering framwork which directly predicts multiple intrinsic properties given a sequence of images. We firstly introduce the problem definition and our notations in Sec 3.1, followed by our model architecture and design choices in Sec 3.2. In Sec 3.3, we introduce a consistency finetuning scheme, improving the model's generalization capability in real-world scenes.

3.1. Problem Definition and Notation

Given a sequence of N images $S = (I_1, ..., I_N), I_i \in \mathbb{R}^{H\times W\times 3}$, the network f aims to map the sequence to a corresponding set of intrinsic images:

$$f((I_i)_{i=1}^N) = (A_i, M_i, R_i, N_i, D_i)_{i=1}^N, \tag{1}$$

where $A_i \in [0,1]^{H\times W\times 3}$ denotes the albedo map, representing the surface's base reflectance color. $M_i \in [0,1]^{H\times W}$ is the metallic map, and $R_i \in [0,1]^{H\times W}$ represents the roughness map. $N_i \in [-1,1]^{H\times W\times 3}$ is the surface normal map, defined in the camera coordinate system. $D_i \in [0,1]^{H\times W\times 3}$ denotes the diffuse shading map, capturing the total incoming radiance over the upper hemisphere at each surface point, which contributes to the diffuse shading.

3.2. MVInverse

Feature Backbone Following recent advances in feed-forward 3D reconstruction [90, 93], we adopt a similar architecture tailored for the multi-view inverse rendering setting, as illustrated in Figure 3.

We use DINOv2 [73] as the visual feature encoder due to its strong generalization capability and ability to extract semantically meaningful high-level representations. Each input image I_i is converted into a set of visual tokens: $X_i =$

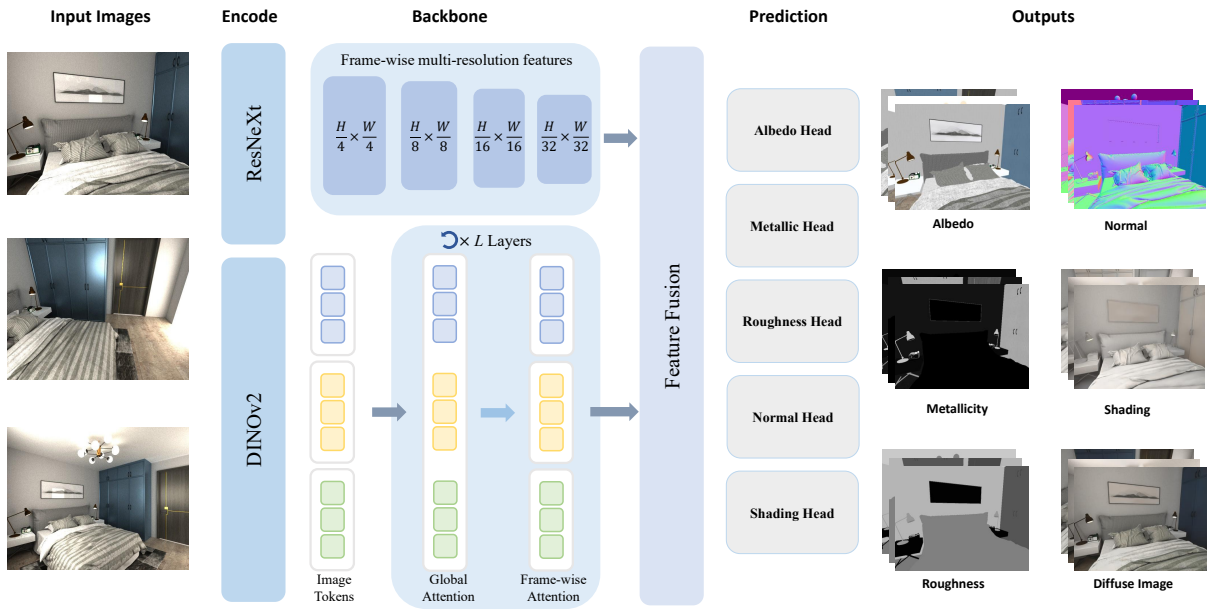

Figure 3. **Framework of MVInverse.** Given an input image sequence, our framework first encodes each frame into patch tokens via DINOv2 [73], where alternating global–frame attention enables cross-view feature aggregation. Concurrently, a frame-wise ResNeXt [99] encoder provides multi-resolution convolutional features to preserve fine spatial details. Notably, these convolutional features are integrated into the transformer stream *only at the final decoder stage*. This DPT-style prediction head produces pixel-aligned intrinsic maps, including albedo, metallic, roughness, normal, and shading. The diffuse image is obtained as the product of albedo and diffuse shading. Further details can be found in the supplementary material.

$\text{DINOv2}(I_i), \quad X_i \in \mathbb{R}^{T\times C}$, where T denotes the number of tokens and C is the feature dimensionality.

To jointly exploit intra-view appearance cues and cross-view intrinsic consistency, we adopt a permutation-equivariant alternating-attention transformer backbone similar to [90, 93]. The backbone alternates between two complementary attention operations:

- **Frame-wise self-attention**, which performs attention *within a single image* and refines token representations by capturing long-range spatial dependencies and local semantic structures.
- **Global self-attention**, which performs attention across all input images and allows tokens from different views to mutually reference and reinforce each other. This mechanism implicitly associates tokens that observe the same underlying 3D surface region, without requiring camera pose supervision.

Through progressive alternation between these two attention operations, the network injects rich intra-view contextual reasoning and globally consistent cross-view information into each token. This yields feature representations that are structurally coherent and view-consistent, forming a reliable basis for multi-view inverse rendering.

Intrinsic Heads Given the rich multi-view latent representation, MVInverse predicts intrinsic properties for every input image through five dedicated prediction heads:

$$P_{i,k} = \text{Head}_k(X_i), \quad k \in \{A, M, R, N, D\}$$

Each head is implemented as a DPT-style dense prediction module [77], producing pixel-aligned predictions at the input resolution.

However, unlike depth and semantic segmentation tasks, where DPT is commonly used, inverse rendering requires much stronger spatial detail preservation. Empirically, we find that using the DPT head alone often leads to overly smoothed and blurry material maps, particularly in albedo, where high-frequency texture patterns must be recovered accurately. To overcome this limitation, we introduce an *auxiliary multi-resolution convolutional encoder* that extracts local high-frequency structures from the input image and injects them into the prediction head via feature fusion. The extracted fine-grained details are skip-connected to the prediction heads to preserve texture sharpness while avoiding unnecessary blurring. As shown in Figure 7, this DPT-hybrid-like [77] architecture substantially improves spatial fidelity and restores fine-scale reflectance variations.

Each intrinsic property is supervised with a tailored loss. For albedo, metallic, roughness, and diffuse shading, we

follow [16, 52, 54] to use mean-squared error (MSE) loss and multi-scale gradient (MSG) loss:

$$\mathcal{L}_{mse}(P) = \frac{1}{N}\sum_{i=1}^{N}(P_i - P_i^*)^2, \tag{2}$$

$$\mathcal{L}_{msg}(P) = \frac{1}{NM}\sum_{i=1}^{N}\sum_{l=1}^{M}\left(\nabla P_{i,l} - \nabla P_{i,l}^*\right)^2, \tag{3}$$

where $P \in \{A, M, R, D\}$ represents the predicted intrinsic map (albedo, metallic, roughness, or diffuse shading), P^* is the corresponding ground-truth supervision, and $\nabla P_{i,l}$ denotes the spatial gradient of P at scale l. Note that a scale-invariant MSE loss is applied for albedo prediction, with details provided in the supplementary materials.

Surface normals are trained using a cosine similarity loss, enforcing correct orientation rather than raw vector magnitude:

$$\mathcal{L}_{normal}(N_i) = 1 - \langle \hat{N}_i, N_i \rangle. \tag{4}$$

Training Our model is trained in an end-to-end manner by optimizing a composite loss function that simultaneously constrains multiple material properties, including albedo, metallicity, roughness, normals, and shading. Each individual loss term is weighted in the overall objective, enabling the model to learn these physical attributes jointly.

$$\begin{aligned}\mathcal{L} = \lambda_{\text{albedo}}\,\mathcal{L}_{\text{albedo}} + \lambda_{\text{metallic}}\,\mathcal{L}_{\text{metallic}} + \lambda_{\text{roughness}}\,\mathcal{L}_{\text{roughness}} \\ + \lambda_{\text{normal}}\,\mathcal{L}_{\text{normal}} + \lambda_{\text{shading}}\,\mathcal{L}_{\text{shading}}\end{aligned} \tag{5}$$

where λ_* denotes the weighting factor for each component, and $\mathcal{L}_*$ corresponds to the loss of the respective material property. In practice, we set all weighting factors $\lambda_* = 1$.

Our model is trained on a diverse collection of datasets to ensure robustness across various environments. Specifically, we use Hypersim [79], Interiorverse [122], PRID [92], Structured3D [120], Amazon-Berkeley Objects [25], MatrixCity [51], and CGIntrinsics [52] as the primary training sources. Since most of these datasets contain synthetic indoor or object-centric scenes, limiting generalization to outdoor scenes, we additionally use DiffusionRenderer [60] to generate pseudo albedo labels for 1,118 real-world videos from the Sekai [58]-Drone dataset. Additional training details and dataset specifications can be found in the supplementary material.

3.3. Consistency Finetuning

After pretraining, the network already produces view-consistent material predictions in synthetic scenes. However, when applied to real-world videos, we observe noticeable temporal flickering on the same 3D regions, resulting in poor visual stability. We resort the issue to the lack of ground-truth supervision in real-world scenes.

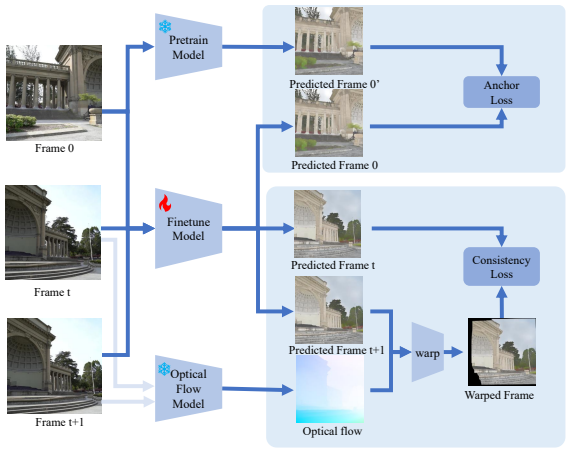

Figure 4. **Consistency finetuning on real videos.** Frames are fed into the finetuned model to produce predictions. Optical flow warps the prediction of frame $t+1$ to frame t, and a consistency loss enforces temporal stability. An anchor loss on frame 0, using a pretrained model's prediction as reference, prevents solution collapse.

To address this, we adopt a two-stage training scheme. After large-scale pretraining on synthetic data, we perform self-supervised finetuning on real-world videos to enhance temporal consistency, as illustrated in Figure 4. Given a sequence of input frames (I_0, I_t, I_{t+1}), the finetuned model generates material predictions $(\hat{M}_0, \hat{M}_t, \hat{M}_{t+1})$. To enforce temporal coherence, we estimate dense optical flow $F_{t+1\to t}$ between frames I_{t+1} and I_t using an off-the-shelf optical flow network [88]. The predicted material map $\hat{M}_{t+1}$ is warped according to $F_{t+1\to t}$ to obtain $\hat{M}^{warp}_{t+1\to t}$. A consistency loss is computed between $\hat{M}^{warp}_{t+1\to t}$ and $\hat{M}_t$, encouraging adjacent predictions to remain consistent under motion.

However, applying only consistency supervision may lead to trivial or collapsed solutions where the model learns to generate temporally smooth but semantically meaningless outputs. To prevent this, we introduce an anchor loss at frame 0. A pretrained model produces a reference material map $\hat{M}_0^{pret}$, and the finetuned model is required to match this prediction at the same frame. This anchor loss preserves the pretrained model's predictive ability while still allowing finetuning to improve temporal stability.

$$\mathcal{L}_{\text{finetune}} = \lambda_{\text{anchor}}\,\|\hat{M}_0 - \hat{M}_0^{\text{pred}}\|_2^2 + \|\hat{M}_t - \hat{M}^{\text{warp}}_{t+1\to t}\|_2^2. \tag{6}$$

Here, we set $\lambda_{\text{anchor}} = 0.1$ to balance the contributions of the anchor and consistency losses. Together, these losses guide the model to learn video-stable material predictions without sacrificing the fidelity obtained from pretraining.

Table 1. **Quantitative evaluation of single-view inverse rendering on Interiorverse [122].** * We do inference in the image mode. Best / 2nd / 3rd per column are highlighted.

	Albedo			Metallic	Roughness
	PSNR ↑	SSIM ↑	LPIPS ↓	RMSE ↓	RMSE ↓
IIW [7]	9.7	0.62	0.47	–	–
RGB↔X [111]	16.4	0.78	0.19	0.44	0.38
ColorfulDiffuse [16]	15.2	0.76	0.35	–	–
IntrinsicImageDiffusion [48]	17.4	0.80	0.22	0.21	0.26
DiffusionRenderer [60]*	21.9	0.87	0.17	0.28	0.35
Ours	23.0	0.92	0.09	0.14	0.17

4. Experiments

We conduct extensive experiments to evaluate the proposed feed-forward multi-view inverse rendering framework. Our evaluation focuses on three core aspects: (1) single-view material estimation quality, (2) multi-view material consistency, and (3) normal estimation quality. Through both quantitative and qualitative comparisons, we demonstrate that our method achieves efficient and consistent scene-level inverse rendering without per-scene optimization.

4.1. Single-view Material Estimation

We first evaluate single-view material estimation on the test set of Interiorverse [122] with 2672 image-material pairs. We evaluate albedo predictions using PSNR, SSIM, and LPIPS, and roughness/metallic using RMSE. For comparison, we include state-of-the-art intrinsic decomposition and inverse rendering methods: DiffusionRenderer [60], ColorfulDiffuse [16], RGB↔X [111], IntrinsicImageDiffusion [48], and earlier approaches [7]. Quantitative results in Table 1 show that our method outperforms all baselines across all evaluated metrics. This demonstrates that our approach not only achieves higher accuracy in albedo and material property estimation but also produces more perceptually faithful and visually consistent results

We further provide qualitative comparisons for zero-shot real-world single-view albedo estimation in Intrinsic Images in the Wild (IIW) [7] dataset, as shown in Figure 5. In the top row, regions highlighted in green indicate areas where RGB↔X [111] and ColorfulDiffuse [16] fail to correctly remove lighting effects, while blue regions show that these methods do not remove the cat's shadow. Additionally, orange regions illustrate that DiffusionRenderer [60] generates unnatural color artifacts, such as a red patch in the leaves, and exhibits suboptimal overall color tone. In the bottom row, red and blue regions correspond to small shadows under the cabinet that are not removed by existing methods. Green regions highlight the shadows of pendant light on the cabinet that other models fail to recognize. In contrast, our method effectively eliminates these shadows, demonstrating superior capability in both shadow removal and albedo recovery under challenging lighting conditions.

Table 2. RMSE for **Albedo**, **Metallic**, and **Roughness**. Results are evaluated on multi-view images in Hypersim [79].

Method	RMSE↓		
	Albedo	Metallic	Roughness
RGB↔X [111]	0.1317	0.3451	0.1813
ColorfulDiffuse [16]	0.0816	-	-
IntrinsicImageDiffusion [48]	0.0878	0.0734	0.0810
DiffusionRenderer [60]	0.0935	0.1250	0.1055
Ours	0.0660	0.0634	0.0259

4.2. Multi-view Material Consistency

In this section, we compare our model's multi-view consistency with existing methods. Since there is no established protocol for evaluating multi-view material consistency, we design our own evaluation scheme. We select the last 50 scenes from the Hypersim [79] test split. Note thta none of the compared methods have seen these five test scenes during training. For each scene, we perform inference on the first 10 views in all camera trajectories. Using known depth maps and camera poses, predicted material maps (albedo, roughness, metallic) from one view are back-projected into 3D world coordinates and reprojected onto other views. We then compute RMSE between reprojected and directly predicted maps over overlapping regions to quantify cross-view consistency. Results in Table 2 show that our method achieves the lowest RMSE across all material channels (albedo, roughness, and metallic), indicating superior cross-view consistency compared to existing approaches, which is an essential property for material acquisition in downstream applications.

We additionally provide qualitative comparisons on multi-view albedo consistency, as shown in Figure 6. We select two views from the *Room* scene in the Mip-NeRF 360 [5] dataset and *Drjohnson* scene in Deep Blending [40] dataset and predict their albedo using each method. The second view is warped to the first view using camera poses and depth estimated by Pi3 [93]. We then compute and visualize an error map between the predicted albedo of the first view and the warped albedo from the second view, where brighter pixels indicate larger inconsistencies between views.

As shown in Figure 6, our model produces the least amount of bright pixels in the error map, indicating significantly lower cross-view discrepancies. This demonstrates that our method maintains strong multi-view albedo consistency, while other approaches exhibit noticeable view-dependent variations.

4.3. Normal Estimation

We evaluate zero-shot surface normal prediction on standard benchmarks, including NYUv2 [86], ScanNet [26], iBims-1 [47], Sintel [12], and OASIS [19]. Normal estimation performance is measured using mean angular er-

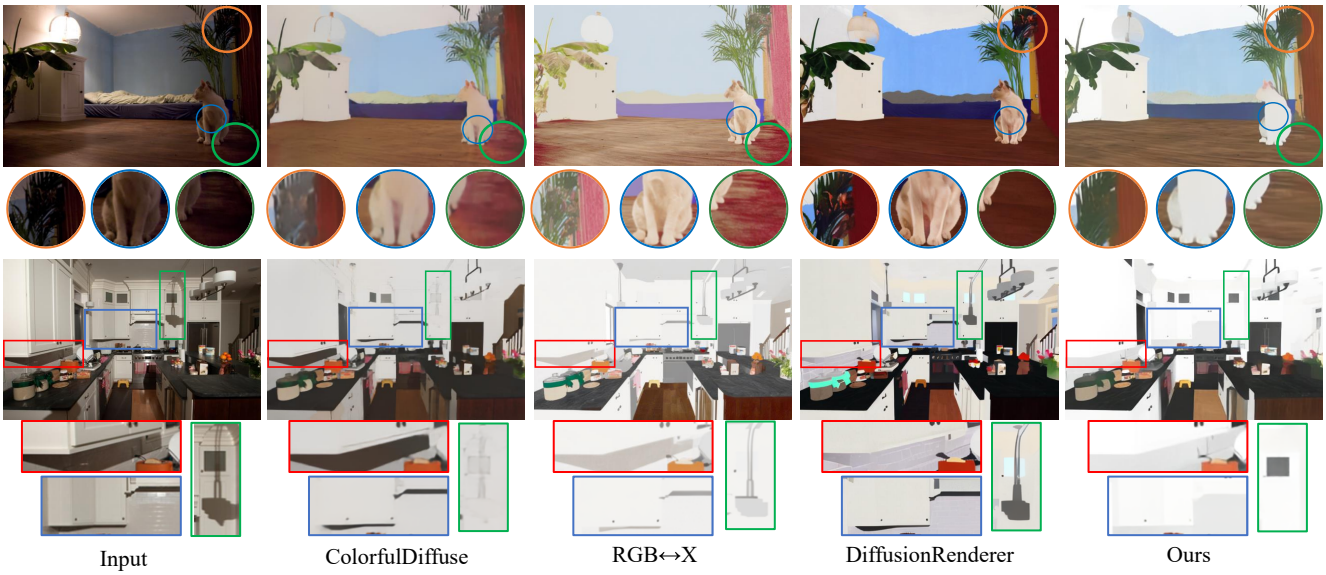

Figure 5. **Single-view Albedo Comparison.** In the top row, green circles highlight regions where lighting is not removed; blue circles highlights regions where shadow still exists; orange circles demonstrate an unnatural red patch inside the leaves. In the bottom row, red and blue regions indicate areas where small shadows are not removed; green regions highlight shadows cast by the pendant light on the cabinet that are not recognized by other models. In contrast, our model handles these cases well.

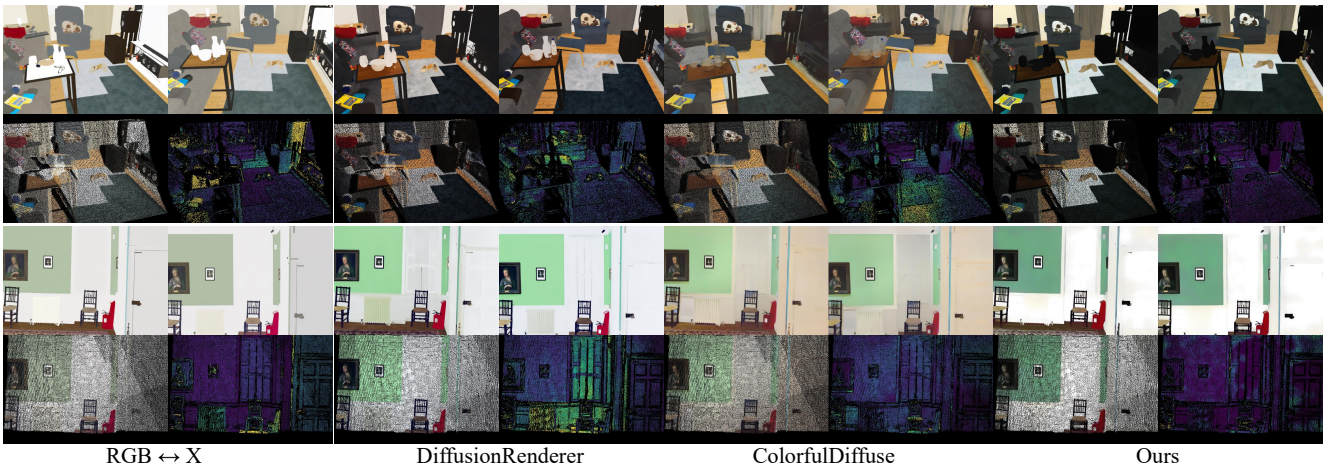

Figure 6. **Multi-view Albedo Comparison** We compare our method with existing approaches under multi-view settings. For each scene, the top-left and top-right subfigures show the albedo predictions from *view 1* and *view 2*, respectively. The bottom-left subfigure shows the result of warping the prediction from *view 2* to *view 1*, and the bottom-right subfigure visualizes the per-pixel difference between the two views. Our method demonstrates higher cross-view consistency and fewer artifacts across viewpoints.

ror (MAE) and the percentage of pixels with angular errors below 11.25° and 30°. We compare against representative state-of-the-art methods, including Omnidata V2 [43], DSINE [3], GeoWizard [32], StableNormal [106], Diffusion-E2E-FT [35], and Lotus-D [39].

Results in Table 3 show that our method achieves the best performance on the majority of datasets and evaluation metrics. In particular, our method achieves the lowest mean angular error and the highest accuracy at 30° thresholds on four out of five benchmarks. Even in the few cases where another model slightly surpasses ours, our performance remains among the top three. This demonstrates that our model generalizes robustly across diverse benchmarks.

4.4. Ablations

Auxiliary multi-resolution encoder preserves fine-grained spatial details. As introduced in Section 3.2, our DPT-style intrinsic heads are augmented with an auxiliary multi-resolution encoder to better preserve high-frequency spatial details. This design allows the network to incorporate fine-grained structural cues from the input image without altering the overall architecture. As shown in Fig. 7, in-

Table 3. **Quantitative evaluation on zero-shot surface normal estimation.** Comparison across five benchmarks. Best / 2nd / 3rd per column are highlighted from deep to light orange.

Method	NYUv2 [86]			ScanNet [26]			iBims-1 [47]			Sintel [12]			OASIS [19]		
	mean ↓	11.25° ↑	30° ↑	mean ↓	11.25° ↑	30° ↑	mean ↓	11.25° ↑	30° ↑	mean ↓	11.25° ↑	30° ↑	mean ↓	11.25° ↑	30° ↑
Omnidata V2 [43]	17.2	55.5	83.0	16.2	60.2	84.7	18.2	63.9	81.1	40.5	14.7	43.5	24.2	27.7	74.2
DSINE [3]	16.4	59.6	83.5	16.2	61.0	84.4	17.1	67.4	82.3	34.9	21.5	52.7	24.4	28.8	72.0
GeoWizard [32]	18.9	50.7	81.5	17.4	53.8	83.5	19.3	63.0	80.3	40.3	12.3	43.5	25.2	23.4	68.1
StableNormal [106]	18.6	53.5	81.7	17.1	57.4	84.1	18.2	65.0	82.4	36.7	14.1	50.7	26.5	23.5	68.7
Diffusion-E2E-FT [35]	16.5	60.4	83.1	14.7	66.1	85.1	16.1	69.7	83.9	33.5	22.3	53.5	23.2	29.4	74.5
Lotus-D [39]	16.2	59.8	83.9	14.7	64.0	86.1	17.1	66.4	83.0	32.3	22.4	57.0	22.3	31.8	76.1
Ours	16.1	60.2	84.1	14.2	66.3	87.1	16.0	69.2	84.1	31.3	22.9	60.2	23.1	29.6	74.5

Figure 7. **Effect of Auxiliary Multi-resolution Encoder.** Incorporating multi-resolution features from the auxiliary encoder helps recover finer details, yielding more accurate and spatially coherent albedo predictions.

corporating the auxiliary encoder leads to more fine-grained and spatially coherent albedo predictions, highlighting the importance of explicitly modeling local details for accurate material recovery.

Consistency finetuning mitigates flickering artifacts. As described in Sec. 3.3, we introduce a self-supervised finetuning strategy to enhance temporal stability on real-world videos. To quantitatively assess its effectiveness, we evaluate the model before and after finetuning on 100 test videos in the DL3DV [64] dataset by measuring the reprojection error across adjacent frames. As shown in Table 4, consistency finetuning notably reduces temporal flickering and yields smoother, more stable material predictions over time.

4.5. Application

Relighting. As in Fig.1, our model predicts view-consistent materials and normals to enable efficient video relighting. By integrating depth and poses from [93], we reconstruct a materialized 3D point cloud for real-time PBR under novel lighting. Our feed-forward framework runs in seconds, demonstrating superior efficiency and interactivity compared to optimization-based methods.

Table 4. **Effect of consistency finetuning.** Comparison before and after applying the proposed self-supervised finetuning on real-world videos.

Method	RMSE↓			
	Albedo	Metallic	Roughness	Shading
Ours (before finetune)	0.0341	0.0250	0.0119	0.0150
Ours (after finetune)	0.0161	0.0146	0.0084	0.0140

Figure 8. **View-Consistent Material Editing**. Our method enables consistent and photorealistic material editing across multiple views. Notably, the shadows and shading on the horn are naturally preserved after editing.

View-consistent Image Editing. With the reconstructed materialized 3D point cloud, we show that we can also perform efficient view-consistent and lighting-aware material editing. Given the region to be modified, we directly replace the albedo of the target points with desired texture and re-render the scene using our predicted diffuse shading, as shown in Figure 8.

5. Conclusion

In this work, we introduced **MVInverse**, a feed-forward multi-view inverse rendering framework that achieves fast and consistent inverse rendering across input views. By synergizing alternating attention with multi-resolution convolutional features, our method bridges the gap between efficiency and spatial fidelity in material recovery. Despite its effectiveness, our approach is still constrained by the limited availability and diversity of material annotations, restricting the model's ability to generalize across diverse scenes. We believe that combining our pipeline with more datasets has the potential to turn feed-forward multi-view inverse rendering into a fundamental component of various downstream tasks.

6. Acknowledgments

This work was supported by the STI 2030-Major Projects under Grant 2021ZD0201404 and Shenzhen Key Laboratory of New Generation Interactive Media Technology Innovation(ZDSYS20210623092001004).

References

[1] Benjamin Attal, Dor Verbin, Ben Mildenhall, Peter Hedman, Jonathan T Barron, Matthew O'Toole, and Pratul P Srinivasan. Flash cache: Reducing bias in radiance cache based inverse rendering. In *European Conference on Computer Vision*, pages 20–36. Springer, 2024. 3

[2] Dejan Azinovic, Tzu-Mao Li, Anton Kaplanyan, and Matthias Nießner. Inverse path tracing for joint material and lighting estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2447–2456, 2019. 2, 3

[3] Gwangbin Bae and Andrew J Davison. Rethinking inductive biases for surface normal estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9535–9545, 2024. 7, 8

[4] Jonathan T Barron and Jitendra Malik. Shape, illumination, and reflectance from shading. *IEEE transactions on pattern analysis and machine intelligence*, 37(8):1670–1687, 2014. 3

[5] Jonathan T. Barron, Ben Mildenhall, Dor Verbin, Pratul P. Srinivasan, and Peter Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. *CVPR*, 2022. 6

[6] Harry Barrow, J Tenenbaum, A Hanson, and E Riseman. Recovering intrinsic scene characteristics. *Comput. vis. syst*, 2(3-26):2, 1978. 3

[7] Sean Bell, Kavita Bala, and Noah Snavely. Intrinsic images in the wild. *ACM Transactions on Graphics (TOG)*, 33(4): 1–12, 2014. 3, 6

[8] Zoubin Bi, Yixin Zeng, Chong Zeng, Fan Pei, Xiang Feng, Kun Zhou, and Hongzhi Wu. Gs3: Efficient relighting with triple gaussian splatting. In *SIGGRAPH Asia 2024 Conference Papers*, pages 1–12, 2024. 3

[9] Mark Boss, Raphael Braun, Varun Jampani, Jonathan T Barron, Ce Liu, and Hendrik Lensch. Nerd: Neural reflectance decomposition from image collections. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12684–12694, 2021. 2, 3

[10] Mark Boss, Varun Jampani, Raphael Braun, Ce Liu, Jonathan Barron, and Hendrik Lensch. Neural-pil: Neural pre-integrated lighting for reflectance decomposition. *Advances in Neural Information Processing Systems*, 34: 10691–10704, 2021.

[11] Mark Boss, Andreas Engelhardt, Abhishek Kar, Yuanzhen Li, Deqing Sun, Jonathan Barron, Hendrik Lensch, and Varun Jampani. Samurai: Shape and material from unconstrained real-world arbitrary image collections. *Advances in Neural Information Processing Systems*, 35:26389–26403, 2022. 3

[12] Daniel J Butler, Jonas Wulff, Garrett B Stanley, and Michael J Black. A naturalistic open source movie for optical flow evaluation. In *European conference on computer vision*, pages 611–625. Springer, 2012. 6, 8

[13] Guangyan Cai, Fujun Luan, Miloš Hašan, Kai Zhang, Sai Bi, Zexiang Xu, Iliyan Georgiev, and Shuang Zhao. Pbirnie: Glossy object capture under non-distant lighting. *arXiv preprint arXiv:2408.06878*, 2024. 3

[14] Junhao Cai, Yuji Yang, Weihao Yuan, Yisheng He, Zilong Dong, Liefeng Bo, Hui Cheng, and Qifeng Chen. Gic: Gaussian-informed continuum for physical property identification and simulation. *Advances in Neural Information Processing Systems*, 37:75035–75063, 2024. 1

[15] Chris Careaga and Yağız Aksoy. Intrinsic image decomposition via ordinal shading. *ACM Trans. Graph.*, 43(1), 2023. 3

[16] Chris Careaga and Yağız Aksoy. Colorful diffuse intrinsic image decomposition in the wild. *ACM Trans. Graph.*, 43 (6), 2024. 2, 3, 5, 6

[17] Chris Careaga and Yağız Aksoy. Physically controllable relighting of photographs. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, pages 1–10, 2025. 1

[18] Hongze Chen, Zehong Lin, and Jun Zhang. Gi-gs: Global illumination decomposition on gaussian splatting for inverse rendering. *arXiv preprint arXiv:2410.02619*, 2024. 3

[19] Weifeng Chen, Shengyi Qian, David Fan, Noriyuki Kojima, Max Hamilton, and Jia Deng. Oasis: A large-scale dataset for single image 3d in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 679–688, 2020. 6, 8

[20] 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. In *European Conference on Computer Vision*, pages 450–467. Springer, 2024. 3

[21] Ziyu Chen, Chenjing Ding, Jianfei Guo, Dongliang Wang, Yikang Li, Xuan Xiao, Wei Wu, and Li Song. L-tracing: Fast light visibility estimation on neural surfaces by sphere tracing. In *European Conference on Computer Vision*, pages 217–233. Springer, 2022. 3

[22] JunYong Choi, SeokYeong Lee, Haesol Park, Seung-Won Jung, Ig-Jae Kim, and Junghyun Cho. Mair: multi-view attention inverse rendering with 3d spatially-varying lighting estimation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8392–8401. IEEE, 2023. 3

[23] JunYong Choi, SeokYeong Lee, Haesol Park, Seung-Won Jung, Ig-Jae Kim, and Junghyun Cho. Mair++: Improving multi-view attention inverse rendering with implicit lighting representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025. 3

[24] Hoon-Gyu Chung, Seokjun Choi, and Seung-Hwan Baek. Differentiable point-based inverse rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4399–4409, 2024. 3

[25] Jasmine Collins, Shubham Goel, Kenan Deng, Achleshwar Luthra, Leon Xu, Erhan Gundogdu, Xi Zhang, Tomas F Yago Vicente, Thomas Dideriksen, Himanshu Arora, et al. Abo: Dataset and benchmarks for real-world 3d object understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 21126–21136, 2022. 2, 5

[26] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5828–5839, 2017. 6, 8

[27] Jan-Niklas Dihlmann, Arjun Majumdar, Andreas Engelhardt, Raphael Braun, and Hendrik Lensch. Subsurface scattering for gaussian splatting. *Advances in Neural Information Processing Systems*, 37:121765–121789, 2024. 3

[28] Hala Djeghim, Nathan Piasco, Luis Roldão, Moussab Bennehar, Dzmitry Tsishkou, Céline Loscos, and Désiré Sidibé. Sail: Self-supervised albedo estimation from real images with a latent diffusion model. *arXiv preprint arXiv:2505.19751*, 2025. 3

[29] Xiaodan Du, Nicholas Kolkin, Greg Shakhnarovich, and Anand Bhattad. Generative models: What do they know? do they know things? let's find out! *arXiv preprint arXiv:2311.17137*, 2023. 3

[30] Xianze Fang, Jingnan Gao, Zhe Wang, Zhuo Chen, Xingyu Ren, Jiangjing Lyu, Qiaomu Ren, Zhonglei Yang, Xiaokang Yang, Yichao Yan, et al. Dens3r: A foundation model for 3d geometry prediction. *arXiv preprint arXiv:2507.16290*, 2025. 3

[31] Graham D Finlayson, Mark S Drew, and Cheng Lu. Intrinsic images by entropy minimization. In *European conference on computer vision*, pages 582–595. Springer, 2004. 3

[32] 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 *European Conference on Computer Vision*, pages 241–258. Springer, 2024. 7, 8

[33] Jian Gao, Chun Gu, Youtian Lin, Zhihao Li, Hao Zhu, Xun Cao, Li Zhang, and Yao Yao. Relightable 3d gaussians: Realistic point cloud relighting with brdf decomposition and ray tracing. In *European Conference on Computer Vision*, pages 73–89. Springer, 2024. 2, 3

[34] Jingnan Gao, Zhe Wang, Xianze Fang, Xingyu Ren, Zhuo Chen, Shengqi Liu, Yuhao Cheng, Jiangjing Lyu, Xiaokang Yang, and Yichao Yan. More: 3d visual geometry reconstruction meets mixture-of-experts. *arXiv preprint arXiv:2510.27234*, 2025. 3

[35] Gonzalo Martin Garcia, Karim Abou Zeid, Christian Schmidt, Daan De Geus, Alexander Hermans, and Bastian Leibe. Fine-tuning image-conditional diffusion models is easier than you think. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 753–762. IEEE, 2025. 7, 8

[36] Roger Grosse, Micah K Johnson, Edward H Adelson, and William T Freeman. Ground truth dataset and baseline evaluations for intrinsic image algorithms. In *2009 IEEE 12th International Conference on Computer Vision*, pages 2335–2342. IEEE, 2009. 3

[37] Chun Gu, Xiaofei Wei, Zixuan Zeng, Yuxuan Yao, and Li Zhang. Irgs: Inter-reflective gaussian splatting with 2d gaussian ray tracing. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 10943–10952, 2025. 3

[38] Yijia Guo, Yuanxi Bai, Liwen Hu, Ziyi Guo, Mianzhi Liu, Yu Cai, Tiejun Huang, and Lei Ma. Prtgs: Precomputed radiance transfer of gaussian splats for real-time high-quality relighting. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 5112–5120, 2024. 3

[39] Jing He, Haodong Li, Wei Yin, Yixun Liang, Leheng Li, Kaiqiang Zhou, Hongbo Zhang, Bingbing Liu, and Ying-Cong Chen. Lotus: Diffusion-based visual foundation model for high-quality dense prediction. *arXiv preprint arXiv:2409.18124*, 2024. 7, 8

[40] Peter Hedman, Julien Philip, True Price, Jan-Michael Frahm, George Drettakis, and Gabriel Brostow. Deep blending for free-viewpoint image-based rendering. *ACM Transactions on Graphics (ToG)*, 37(6):1–15, 2018. 6

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

[42] Yingwenqi Jiang, Jiadong Tu, Yuan Liu, Xifeng Gao, Xiaoxiao Long, Wenping Wang, and Yuexin Ma. Gaussianshader: 3d gaussian splatting with shading functions for reflective surfaces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5322–5332, 2024. 2, 3

[43] Oğuzhan Fatih Kar, Teresa Yeo, Andrei Atanov, and Amir Zamir. 3d common corruptions and data augmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18963–18974, 2022. 7, 8

[44] Nikhil Keetha, Norman Müller, Johannes Schönberger, Lorenzo Porzi, Yuchen Zhang, Tobias Fischer, Arno Knapitsch, Duncan Zauss, Ethan Weber, Nelson Antunes, et al. Mapanything: Universal feed-forward metric 3d reconstruction. *arXiv preprint arXiv:2509.13414*, 2025. 2

[45] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023. 3

[46] Julian Knodt, Joe Bartusek, Seung-Hwan Baek, and Felix Heide. Neural ray-tracing: Learning surfaces and reflectance for relighting and view synthesis. *arXiv preprint arXiv:2104.13562*, 2021. 2, 3

[47] Tobias Koch, Lukas Liebel, Friedrich Fraundorfer, and Marco Korner. Evaluation of cnn-based single-image depth estimation methods. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 0–0, 2018. 6, 8

[48] Peter Kocsis, Vincent Sitzmann, and Matthias Nießner. Intrinsic image diffusion for indoor single-view material estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5198–5208, 2024. 2, 3, 6

[49] Shuichang Lai, Letian Huang, Jie Guo, Kai Cheng, Bowen Pan, Xiaoxiao Long, Jiangjing Lyu, Chengfei Lv, and Yanwen Guo. Glossygs: Inverse rendering of glossy objects

with 3d gaussian splatting. *IEEE Transactions on Visualization and Computer Graphics*, 2025. 3

[50] Edwin H Land and John J McCann. Lightness and retinex theory. *Journal of the Optical society of America*, 61(1): 1–11, 1971. 3

[51] Yixuan Li, Lihan Jiang, Linning Xu, Yuanbo Xiangli, Zhenzhi Wang, Dahua Lin, and Bo Dai. Matrixcity: A large-scale city dataset for city-scale neural rendering and beyond. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3205–3215, 2023. 2, 5

[52] Zhengqi Li and Noah Snavely. Cgintrinsics: Better intrinsic image decomposition through physically-based rendering. In *Proceedings of the European conference on computer vision (ECCV)*, pages 371–387, 2018. 2, 5

[53] Zhengqi Li and Noah Snavely. Learning intrinsic image decomposition from watching the world. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9039–9048, 2018. 2, 3

[54] Zhengqi Li and Noah Snavely. Megadepth: Learning single-view depth prediction from internet photos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2041–2050, 2018. 5

[55] Zhengqin Li, Kalyan Sunkavalli, and Manmohan Chandraker. Materials for masses: Svbrdf acquisition with a single mobile phone image. In *Proceedings of the European conference on computer vision (ECCV)*, pages 72–87, 2018. 2, 3

[56] Zhengqin Li, Mohammad Shafiei, Ravi Ramamoorthi, Kalyan Sunkavalli, and Manmohan Chandraker. Inverse rendering for complex indoor scenes: Shape, spatially-varying lighting and svbrdf from a single image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2475–2484, 2020.

[57] Zhengqin Li, Jia Shi, Sai Bi, Rui Zhu, Kalyan Sunkavalli, Miloš Hašan, Zexiang Xu, Ravi Ramamoorthi, and Manmohan Chandraker. Physically-based editing of indoor scene lighting from a single image. In *European Conference on Computer Vision*, pages 555–572. Springer, 2022. 2, 3

[58] Zhen Li, Chuanhao Li, Xiaofeng Mao, Shaoheng Lin, Ming Li, Shitian Zhao, Zhaopan Xu, Xinyue Li, Yukang Feng, Jianwen Sun, et al. Sekai: A video dataset towards world exploration. *arXiv preprint arXiv:2506.15675*, 2025. 5

[59] Ruofan Liang, Zan Gojcic, Merlin Nimier-David, David Acuna, Nandita Vijaykumar, Sanja Fidler, and Zian Wang. Photorealistic object insertion with diffusion-guided inverse rendering. In *European Conference on Computer Vision*, pages 446–465. Springer, 2024. 3

[60] Ruofan Liang, Zan Gojcic, Huan Ling, Jacob Munkberg, Jon Hasselgren, Chih-Hao Lin, Jun Gao, Alexander Keller, Nandita Vijaykumar, Sanja Fidler, et al. Diffusion renderer: Neural inverse and forward rendering with video diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 26069–26080, 2025. 2, 5, 6

[61] Ruofan Liang, Kai He, Zan Gojcic, Igor Gilitschenski, Sanja Fidler, Nandita Vijaykumar, and Zian Wang. Luxdit: Lighting estimation with video diffusion transformer. *arXiv preprint arXiv:2509.03680*, 2025. 3

[62] Zhihao Liang, Qi Zhang, Ying Feng, Ying Shan, and Kui Jia. Gs-ir: 3d gaussian splatting for inverse rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21644–21653, 2024. 2, 3

[63] Chih-Hao Lin, Jia-Bin Huang, Zhengqin Li, Zhao Dong, Christian Richardt, Tuotuo Li, Michael Zollhöfer, Johannes Kopf, Shenlong Wang, and Changil Kim. Iris: Inverse rendering of indoor scenes from low dynamic range images. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 465–474, 2025. 3

[64] Lu Ling, Yichen Sheng, Zhi Tu, Wentian Zhao, Cheng Xin, Kun Wan, Lantao Yu, Qianyu Guo, Zixun Yu, Yawen Lu, et al. Dl3dv-10k: A large-scale scene dataset for deep learning-based 3d vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22160–22169, 2024. 8

[65] Yehonathan Litman, Or Patashnik, Kangle Deng, Aviral Agrawal, Rushikesh Zawar, Fernando De la Torre, and Shubham Tulsiani. Materialfusion: Enhancing inverse rendering with material diffusion priors. In *2025 International Conference on 3D Vision (3DV)*, pages 802–812. IEEE, 2025. 3

[66] Yuan Liu, Peng Wang, Cheng Lin, Xiaoxiao Long, Jiepeng Wang, Lingjie Liu, Taku Komura, and Wenping Wang. Nero: Neural geometry and brdf reconstruction of reflective objects from multiview images. *ACM Transactions on Graphics (ToG)*, 42(4):1–22, 2023. 3

[67] Yifan Liu, Zhiyuan Min, Zhenwei Wang, Junta Wu, Tengfei Wang, Yixuan Yuan, Yawei Luo, and Chunchao Guo. Worldmirror: Universal 3d world reconstruction with any-prior prompting. *arXiv preprint arXiv:2510.10726*, 2025. 2

[68] Jiahao Lu, Tianyu Huang, Peng Li, Zhiyang Dou, Cheng Lin, Zhiming Cui, Zhen Dong, Sai-Kit Yeung, Wenping Wang, and Yuan Liu. Align3r: Aligned monocular depth estimation for dynamic videos. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 22820–22830, 2025. 3

[69] 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 (TOG)*, 42(6):1–14, 2023. 3

[70] Wei-Chiu Ma, Hang Chu, Bolei Zhou, Raquel Urtasun, and Antonio Torralba. Single image intrinsic decomposition without a single intrinsic image. In *Proceedings of the European conference on computer vision (ECCV)*, pages 201–217, 2018. 3

[71] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 3

[72] Merlin Nimier-David, Zhao Dong, Wenzel Jakob, Anton Kaplanyan, Adrien Bousseau, and Morgan McGuire. Ma-

terial and lighting reconstruction for complex indoor scenes with texture-space differentiable rendering. In *EGSR (DL)*, pages 73–84, 2021. 2, 3

[73] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 3, 4

[74] Weikun Peng, Sota Taira, Chris Careaga, and Yağız Aksoy. Interactive object insertion with differentiable rendering. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Posters*, pages 1–3, 2025. 1

[75] Julien Philip, Sébastien Morgenthaler, Michaël Gharbi, and George Drettakis. Free-viewpoint indoor neural relighting from multi-view stereo. *ACM Transactions on Graphics (TOG)*, 40(5):1–18, 2021. 3

[76] Pakkapon Phongthawee, Worameth Chinchuthakun, Nontaphat Sinsunthithet, Varun Jampani, Amit Raj, Pramook Khungurn, and Supasorn Suwajanakorn. Diffusionlight: Light probes for free by painting a chrome ball. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 98–108, 2024. 3

[77] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 12179–12188, 2021. 4

[78] Kerui Ren, Jiayang Bai, Linning Xu, Lihan Jiang, Jiangmiao Pang, Mulin Yu, and Bo Dai. Mv-colight: Efficient object compositing with consistent lighting and shadow generation. *arXiv preprint arXiv:2505.21483*, 2025. 1

[79] Mike Roberts, Jason Ramapuram, Anurag Ranjan, Atulit Kumar, Miguel Angel Bautista, Nathan Paczan, Russ Webb, and Joshua M Susskind. Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10912–10922, 2021. 2, 5, 6

[80] Shen Sang and Manmohan Chandraker. Single-shot neural relighting and svbrdf estimation. In *European Conference on Computer Vision*, pages 85–101. Springer, 2020. 3

[81] Soumyadip Sengupta, Jinwei Gu, Kihwan Kim, Guilin Liu, David W Jacobs, and Jan Kautz. Neural inverse rendering of an indoor scene from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8598–8607, 2019. 3

[82] Jianbing Shen, Xiaoshan Yang, Yunde Jia, and Xuelong Li. Intrinsic images using optimization. In *CVPR 2011*, pages 3481–3487. IEEE, 2011. 3

[83] Li Shen, Ping Tan, and Stephen Lin. Intrinsic image decomposition with non-local texture cues. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–7. IEEE, 2008. 3

[84] Jian Shi, Yue Dong, Hao Su, and Stella X Yu. Learning non-lambertian object intrinsics across shapenet categories. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1685–1694, 2017. 3

[85] Yahao Shi, Yanmin Wu, Chenming Wu, Xing Liu, Chen Zhao, Haocheng Feng, Jian Zhang, Bin Zhou, Errui Ding, and Jingdong Wang. Gir: 3d gaussian inverse rendering for relightable scene factorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025. 3

[86] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *European conference on computer vision*, pages 746–760. Springer, 2012. 6, 8

[87] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7495–7504, 2021. 2, 3

[88] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *European conference on computer vision*, pages 402–419. Springer, 2020. 5

[89] Haoyuan Wang, Wenbo Hu, Lei Zhu, and Rynson WH Lau. Inverse rendering of glossy objects via the neural plenoptic function and radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19999–20008, 2024. 3

[90] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5294–5306, 2025. 2, 3, 4

[91] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. *arXiv preprint arXiv:2106.10689*, 2021. 3

[92] Yujie Wang, Qingnan Fan, Kun Li, Dongdong Chen, Jingyu Yang, Jianzhi Lu, Dani Lischinski, and Baoquan Chen. High quality rendered dataset and non-local graph convolutional network for intrinsic image decomposition. *Journal of Image and Graphics*, 2022. 2, 5

[93] Yifan Wang, Jianjun Zhou, Haoyi Zhu, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Jiangmiao Pang, Chunhua Shen, and Tong He. π^3: Permutation-equivariant visual geometry learning. *arXiv preprint arXiv:2507.13347*, 2025. 2, 3, 4, 6, 8

[94] Zian Wang, Jonah Philion, Sanja Fidler, and Jan Kautz. Learning indoor inverse rendering with 3d spatially-varying lighting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12538–12547, 2021. 3

[95] Felix Wimbauer, Shangzhe Wu, and Christian Rupprecht. De-rendering 3d objects in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18490–18499, 2022. 3

[96] Liwen Wu, Rui Zhu, Mustafa B Yaldiz, Yinhao Zhu, Hong Cai, Janarbek Matai, Fatih Porikli, Tzu-Mao Li, Manmohan Chandraker, and Ravi Ramamoorthi. Factorized inverse path tracing for efficient and accurate material-lighting estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3848–3858, 2023. 2, 3

[97] Tong Wu, Jia-Mu Sun, Yu-Kun Lai, Yuewen Ma, Leif Kobbelt, and Lin Gao. Deferredgs: Decoupled and editable

gaussian splatting with deferred shading. *arXiv preprint arXiv:2404.09412*, 2024. 3

[98] Jianfeng Xiang, Zelong Lv, Sicheng Xu, Yu Deng, Ruicheng Wang, Bowen Zhang, Dong Chen, Xin Tong, and Jiaolong Yang. Structured 3d latents for scalable and versatile 3d generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 21469–21480, 2025. 1

[99] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500, 2017. 4

[100] Bojun Xiong, Jialun Liu, Jiakui Hu, Chenming Wu, Jinbo Wu, Xing Liu, Chen Zhao, Errui Ding, and Zhouhui Lian. Texgaussian: Generating high-quality pbr material via octree-based 3d gaussian splatting. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 551–561, 2025. 1

[101] Xinli Xu, Wenhang Ge, Dicong Qiu, ZhiFei Chen, Dongyu Yan, Zhuoyun Liu, Haoyu Zhao, Hanfeng Zhao, Shunsi Zhang, Junwei Liang, et al. Gaussianproperty: Integrating physical properties to 3d gaussians with lmms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7231–7240, 2025. 1

[102] Zunnan Xu, Zhentao Yu, Zixiang Zhou, Jun Zhou, Xiaoyu Jin, Fa-Ting Hong, Xiaozhong Ji, Junwei Zhu, Chengfei Cai, Shiyu Tang, et al. Hunyuanportrait: Implicit condition control for enhanced portrait animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15909–15919, 2025. 3

[103] Wenqi Yang, Guanying Chen, Chaofeng Chen, Zhenfang Chen, and Kwan-Yee K Wong. Ps-nerf: Neural inverse rendering for multi-view photometric stereo. In *European conference on computer vision*, pages 266–284. Springer, 2022. 3

[104] Ziyi Yang, Xinyu Gao, Wu Yu, Xiaowei Zhou, and Xiaogang Jin. Robir: Robust inverse rendering for high-illumination scenes. *Advances in Neural Information Processing Systems*, 37:124114–124136, 2024.

[105] Yao Yao, Jingyang Zhang, Jingbo Liu, Yihang Qu, Tian Fang, David McKinnon, Yanghai Tsin, and Long Quan. Neilf: Neural incident light field for physically-based material estimation. In *European conference on computer vision*, pages 700–716. Springer, 2022. 3

[106] 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)*, 43(6):1–18, 2024. 7, 8

[107] Keyang Ye, Qiming Hou, and Kun Zhou. Progressive radiance distillation for inverse rendering with gaussian splatting. *arXiv preprint arXiv:2408.07595*, 2024. 3

[108] Kai Ye, Chong Gao, Guanbin Li, Wenzheng Chen, and Baoquan Chen. Geosplatting: Towards geometry guided gaussian splatting for physically-based inverse rendering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 28991–29000, 2025. 3

[109] Bohan Yu, Siqi Yang, Xuanning Cui, Siyan Dong, Baoquan Chen, and Boxin Shi. Milo: Multi-bounce inverse rendering for indoor scene with light-emitting objects. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(8):10129–10142, 2023. 2, 3

[110] Ye Yu and William AP Smith. Inverserendernet: Learning single image inverse rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3155–3164, 2019. 3

[111] Zheng Zeng, Valentin Deschaintre, Iliyan Georgiev, Yannick Hold-Geoffroy, Yiwei Hu, Fujun Luan, Ling-Qi Yan, and Miloš Hašan. Rgb ↔ x: Image decomposition and synthesis using material- and lighting-aware diffusion models. In *ACM SIGGRAPH 2024 Conference Papers*, New York, NY, USA, 2024. Association for Computing Machinery. 2, 3, 6

[112] Albert J Zhai, Yuan Shen, Emily Y Chen, Gloria X Wang, Xinlei Wang, Sheng Wang, Kaiyu Guan, and Shenlong Wang. Physical property understanding from language-embedded feature fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 28296–28305, 2024. 1

[113] Fangneng Zhan, Yingchen Yu, Changgong Zhang, Rongliang Wu, Wenbo Hu, Shijian Lu, Feiying Ma, Xuansong Xie, and Ling Shao. Gmlight: Lighting estimation via geometric distribution approximation. *IEEE Transactions on Image Processing*, 31:2268–2278, 2022. 3

[114] Kai Zhang, Fujun Luan, Zhengqi Li, and Noah Snavely. Iron: Inverse rendering by optimizing neural sdfs and materials from photometric images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5565–5574, 2022. 3

[115] Qing Zhang, Jin Zhou, Lei Zhu, Wei Sun, Chunxia Xiao, and Wei-Shi Zheng. Unsupervised intrinsic image decomposition using internal self-similarity cues. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12): 9669–9686, 2021. 3

[116] Qian Zhang, Vikas Thamizharasan, and James Tompkin. Learning physically based material and lighting decompositions for face editing. *Computational Visual Media*, 10 (2):295–308, 2024. 3

[117] Xiuming Zhang, Pratul P Srinivasan, Boyang Deng, Paul Debevec, William T Freeman, and Jonathan T Barron. Nerfactor: Neural factorization of shape and reflectance under an unknown illumination. *ACM Transactions on Graphics (ToG)*, 40(6):1–18, 2021. 2, 3

[118] Qi Zhao, Ping Tan, Qiang Dai, Li Shen, Enhua Wu, and Stephen Lin. A closed-form solution to retinex with nonlocal texture constraints. *IEEE transactions on pattern analysis and machine intelligence*, 34(7):1437–1444, 2012. 3

[119] Zibo Zhao, Zeqiang Lai, Qingxiang Lin, Yunfei Zhao, Haolin Liu, Shuhui Yang, Yifei Feng, Mingxin Yang, Sheng Zhang, Xianghui Yang, et al. Hunyuan3d 2.0: Scaling diffusion models for high resolution textured 3d assets generation. *arXiv preprint arXiv:2501.12202*, 2025. 1

[120] Jia Zheng, Junfei Zhang, Jing Li, Rui Tang, Shenghua Gao, and Zihan Zhou. Structured3d: A large photo-realistic

dataset for structured 3d modeling. In *European Conference on Computer Vision*, pages 519–535. Springer, 2020. 2, 5

[121] Jun Zhou, Jiahao Li, Zunnan Xu, Hanhui Li, Yiji Cheng, Fa-Ting Hong, Qin Lin, Qinglin Lu, and Xiaodan Liang. Fireedit: Fine-grained instruction-based image editing via region-aware vision language model. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 13093–13103, 2025. 3

[122] Jingsen Zhu, Fujun Luan, Yuchi Huo, Zihao Lin, Zhihua Zhong, Dianbing Xi, Rui Wang, Hujun Bao, Jiaxiang Zheng, and Rui Tang. Learning-based inverse rendering of complex indoor scenes with differentiable monte carlo raytracing. In *SIGGRAPH Asia 2022 Conference Papers*, pages 1–8, 2022. 2, 3, 5, 6

[123] Rui Zhu, Zhengqin Li, Janarbek Matai, Fatih Porikli, and Manmohan Chandraker. Irisformer: Dense vision transformers for single-image inverse rendering in indoor scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2822–2831, 2022. 3

[124] Zuo-Liang Zhu, Beibei Wang, and Jian Yang. Gs-ror: 3d gaussian splatting for reflective object relighting via sdf priors. *arXiv e-prints*, pages arXiv–2406, 2024. 3