

Supplementary Material — Intrinsic PAPR: Tackling Misattribution in 3D Intrinsic Decomposition via Proximity Attention Point Rendering

Alireza Moazeni¹^{*}, Shichong Peng¹, Yanshu Zhang¹, Chirag Vashist¹,
and Ke Li^{1,2,3}

¹ APEX Lab, School of Computing Science, Simon Fraser University, Canada

² Alberta Machine Intelligence Institute (Amii), Canada

³ Canadian Institute for Advanced Research (CIFAR), Canada

{sam62, shichong_peng, yanshu_zhang, chirag_vashist, keli}@sfu.ca

This document provides extended implementation details, additional quantitative and qualitative results, and further ablations for Intrinsic PAPR, complementing the main paper.

1	Implementation Details	2
1.1	PAPR Renderer and Intrinsic Adaptation	2
1.2	Ambiguity-aware Albedo Estimates	3
1.3	Data-preprocessing	5
1.4	Computational Cost for Editing	5
1.5	Training Hyperparameters	5
1.6	Training Complexity	5
2	Additional Results	7
2.1	Attention Concentration Analysis	7
2.2	Per-scene Quantitative Evaluation on Synthetic Datasets	9
2.3	Per-scene Quantitative Evaluation on Real-world Dataset	9
2.4	Per-component Ablation: Per-scene Results	9
2.5	Depth and Novel View Synthesis Comparison with PAPR	9
2.6	Evaluating the Impact of Pseudo-Ground Truth Albedo Supervision on Baselines	13
2.7	Albedo Consistency Across Different Views	15
2.8	Decomposition Results on Real-world Scenes	18
2.9	Albedo Transfer	20
2.10	Shading Transfer	23
2.11	Transfer Error vs. Number of Training Views	23
2.12	Scene-level Shading Intensity Control	26
2.13	Custom Editing Tool	27
2.14	Robustness, Limitations, and Future Work	29

^{*} Corresponding author.

1 Implementation Details

1.1 PAPR Renderer and Intrinsic Adaptation

Figure 1 and Figure 2 provide the architectural details of our intrinsic variant of the PAPR renderer [17]. Each 3D point stores two learnable feature vectors, an albedo feature $\mathbf{a}_i$ and a shading feature $\mathbf{h}_i$, together with its position and influence score. For a given camera ray, we follow PAPR and use a proximity-based attention module to select the top- K nearby points, compute a scalar attention logit from the point position, its perpendicular and along-ray distances to the ray, and its influence, and normalize these logits into weights along the ray. We then aggregate $\mathbf{a}_i$ and $\mathbf{h}_i$ separately with the same weights, yielding per-pixel albedo and shading feature maps A_C and S_C as defined in Sec. 3 of the main paper. The decoder U-Net in Figure 2 takes the concatenation $[A_C, S_C]$ as input and predicts the final RGB image; the per-pixel albedo map is decoded separately from the albedo feature map A_C .

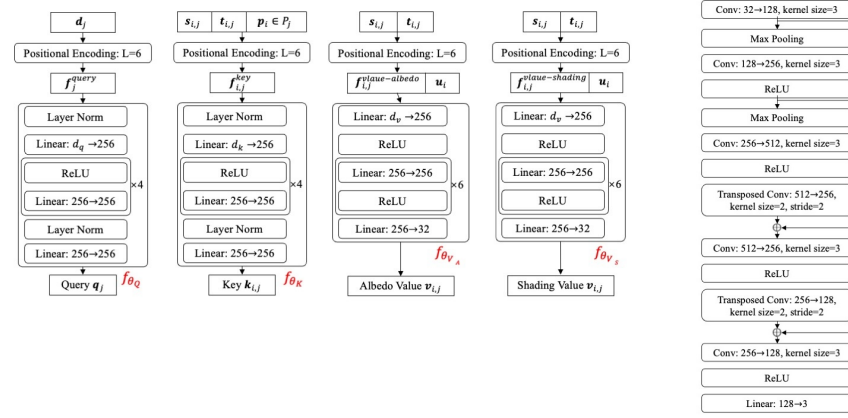


Fig. 1: Embedding MLPs architecture details showing the network structure and connections.

Fig. 2: U-Net architecture details showing the encoder-decoder structure.

1.2 Ambiguity-aware Albedo Estimates

Our method relies on an ambiguity-aware 2D albedo prior to supply per-view hypothesis sets for the space-carving loss. The training and inference of this prior are strictly decoupled from the Intrinsic PAPR optimization pipeline, ensuring that the 2D prior does not introduce any computational or memory overhead during the main 3D training phase. The preparation and integration of the 2D prior follow a three-step offline procedure:

1. Architecture Modification and Fine-tuning. Figure 3 provides the architectural details of our ambiguity-aware 2D albedo prior, which extends the intrinsic network of Careaga et al. [2]. We insert AdaIN layers into the encoder blocks of the original prior and predict their scale and shift parameters from a latent code $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ using a lightweight MLP, so that different samples of $\mathbf{z}$ modulate the encoder features into distinct albedo hypotheses. During training, for each input image we draw M latent codes, obtain a set of candidate albedo maps $\{G(I, z_i)\}_{i=1}^M$, and apply a cIMLE-style objective [8, 12]: we measure the reconstruction error to the pseudo-ground-truth albedo for all candidates and backpropagate only through the best-matching one. This procedure encourages the prior to represent the full range of plausible albedo explanations—capturing ambiguities such as albedo–shading tradeoffs or texture vs. geometry—while remaining faithful to the supervision. This fine-tuning is performed once, independently of any specific scene, and requires approximately 1 hour for 100 steps on a single NVIDIA RTX 3090 GPU using the original training dataset of the prior model.

2. Offline Sample Generation. Using the single fine-tuned model from the previous step, we perform a one-time offline inference for each scene independently of Intrinsic PAPR training. We feed the training views of a given scene into the fine-tuned ambiguity-aware model, generate M candidate samples per view, and save them to disk. This procedure is run once per scene and takes approximately 2 minutes for 100 views on an RTX 3090 GPU. These saved samples constitute the per-view hypothesis set utilized by our space-carving loss in Sec. 3 of the main paper.

3. Integration with Intrinsic PAPR. During the actual Intrinsic PAPR optimization on a 3D scene, we simply load all pre-computed albedo samples directly into memory. Because we do not perform any 2D network inference or maintain the prior’s weights in GPU memory during the 3D phase, our method introduces zero computational time overhead from the 2D prior.

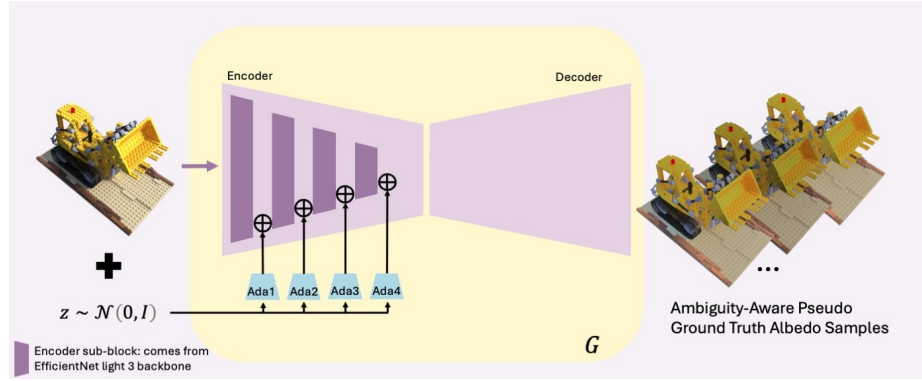


Fig. 3: Architecture of our ambiguity-aware albedo estimation module. The network incorporates AdaIN layers that apply scale and shift based on a latent code $z \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, enabling generation of diverse albedo maps.

1.3 Data-preprocessing

For the ground truth colour and albedo images, we preprocess them to ensure that the intensities in linear space are within the range of $[0,1]$, and then we convert them to the logarithmic domain for improved numerical stability during training.

1.4 Computational Cost for Editing

Editing in our method is performed by directly modifying the learnable point features ($\mathbf{a}_i$ for albedo, $\mathbf{h}_i$ for shading) in the already-optimized scene representation. Given a user-defined region, the features of source points are transferred to target points in a single pre-render pass, with no gradient computation, no network re-training, and no further optimization of any kind. The edited scene is then rendered through the standard PAPR pipeline, which is unchanged. Consequently, an edited scene renders at the same speed as the original, approximately 20 FPS on a single GPU. This stands in contrast to methods such as Seal-3D [13], which require fine-tuning network weights to apply each edit.

1.5 Training Hyperparameters

We set the feature dimensions $m = n = 32$ and $d_{\text{albedo}} = d_{\text{shading}} = 16$, initialize each view’s learnable scalar to 0.85, use $M = 10$ albedo samples per view, and optimize with the Adam optimizer [6]. For all real-world datasets, we set $w_{SC} = 0.125$ and for synthetic datasets, we set the Space Carving loss weight, $w_{SC} = 0.085$.

1.6 Training Complexity

Intrinsic PAPR is primarily designed to advance the accuracy and reliability of 3D intrinsic decomposition; training and inference efficiency were not primary design objectives of this work. That said, the following reports the practical costs of our implementation.

Training. Our method follows the same 250K-iteration training schedule as vanilla PAPR [17], completing in approximately 3 GPU-hours per scene on a single NVIDIA L40S GPU. The integration of the ambiguity-aware 2D prior introduces no per-iteration overhead: as described in Sec. 1, the M albedo hypotheses for all training views are pre-computed offline and loaded into memory prior to training. Consequently, the overall training time of Intrinsic PAPR is comparable to that of the original PAPR renderer.

Inference. At test time, our method uses the same rendering pipeline as vanilla PAPR without any modification. Rendering proceeds at approximately 20 FPS on a single GPU, identical to vanilla PAPR.

Peak memory. GPU memory peaks at 8–15 GB with the default top- $K = 20$ proximity attention and 13–24 GB with a larger top- $K = 64$ neighbourhood, depending on scene complexity.

Comparison to GS-IR. Gaussian-splatting inverse-rendering methods such as GS-IR [9] render faster than our attention-based pipeline, but are far less accurate for point-level editing: our per-primitive albedo-transfer error stays $5.04\text{--}5.23\times$ below GS-IR across 25 to 200 training views (Table 8). We deliberately trade rendering speed for markedly more accurate, directly editable per-primitive intrinsics.

2 Additional Results

2.1 Attention Concentration Analysis

To substantiate the claim that Intrinsic PAPR’s proximity attention leads to highly concentrated supervision—and thereby improved per-point identifiability and reduced misattribution—we quantitatively analyse the attention concentration during inference. For each foreground ray, we compute the maximum normalized attention weight (w_k^{ij}) and the Shannon entropy of the attention distribution over the top- K nearest points in the point cloud.

As shown in Figure 4, the proximity attention is extremely concentrated. The attention entropy histogram has most of its mass at zero (Figure 4b), indicating that the rendering of a pixel is almost entirely determined by a single 3D point for most rays. Correspondingly, a significant majority of foreground rays place around 90% or more of their attention mass on a single point—a maximum attention weight near 1.0 (Figure 4a).

Furthermore, we verify that each foreground point is supervised by a narrow cone of rays by counting, for each point, the number of rays that strongly supervise it—i.e., rays where that point has the highest attention weight and the maximum weight exceeds 0.8. As illustrated in Figure 4c, this distribution is highly concentrated: approximately 90% of the foreground points are strongly supervised by exactly one ray, with the remaining foreground points receiving supervision from only a handful of rays.

This geometric locality is the primary mechanism behind Intrinsic PAPR’s ability to mitigate misattribution. Unlike volumetric splatting or standard NeRFs, where transmittance is distributed multiplicatively along a long ray, our formulation ensures that the photometric residual and albedo supervision backpropagate primarily to a single, specific surface point. This concentrated, point-wise identifiability prevents compensatory effects where multiple fuzzy primitives jointly recreate a pixel colour, resulting in the crisp, view-consistent albedo and shading factorizations demonstrated throughout our experiments.

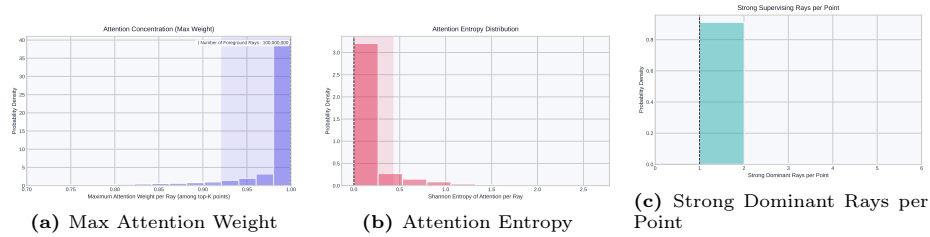


Fig. 4: Attention Concentration Analysis. We quantify the proximity attention weights during the rendering process to demonstrate the geometric locality of Intrinsic PAPR. **(a) Max Attention Weight:** The histogram of maximum attention weights per ray shows a strong peak near 1.0, indicating that a single point strongly dominates the rendered colour of most pixels. **(b) Attention Entropy:** The histogram of attention entropy confirms that the distribution over the top- K points is highly concentrated, with most mass at zero. **(c) Strong Dominant Rays per Point:** For each point, we count rays where it is dominant and the max attention weight ≥ 0.8 . Approximately 90% of foreground points are strongly supervised by exactly one ray, demonstrating that backpropagated supervision is confined to a tight, local cone. Together, these metrics confirm that our proximity-based representation enforces point-wise identifiability, which directly mitigates the misattribution of albedo and shading commonly observed in volumetric scattering methods.

2.2 Per-scene Quantitative Evaluation on Synthetic Datasets

In Table 1, we present a detailed scene-by-scene comparison of novel view reconstruction on synthetic datasets, including NeRF Synthetic [11] and TensoIR [5]. In Table 2, we provide the corresponding per-scene breakdown for the quantitative evaluation of albedo quality on the NeRF Synthetic dataset. Quantitative albedo evaluation is performed on synthetic datasets because real-world captures do not provide albedo ground truth; this follows standard practice in prior work [9, 16]. Following prior work [9, 16], albedo metrics are computed after per-view scale-and-shift alignment, applied uniformly to all compared methods. For a fair, apples-to-apples comparison, we re-run every baseline under this single unified protocol rather than copying their published numbers; reported values therefore differ from the originals wherever the original setups differ from ours—in test splits, scene re-renderings under different environment maps, and per-channel albedo rescaling. Together, these per-scene tables complement the averaged metrics of the main paper by providing a detailed view of how performance varies across individual scenes.

2.3 Per-scene Quantitative Evaluation on Real-world Dataset

In Table 3 we present a per-scene breakdown of novel view reconstruction quality on our real-world benchmarks, comprising the Tanks & Temples subset [7] and the real-scene Mip-NeRF 360 dataset [1].

2.4 Per-component Ablation: Per-scene Results

The main paper reports the component ablation averaged over the synthetic scenes; Table 4 gives the per-scene breakdown for all three metrics. The ordering is consistent across scenes—the full model attains the best NVS PSNR, albedo PSNR, and MACE on every scene—and the largest NVS gains appear on the most geometrically and texturally complex scenes, where the 2D prior is least reliable and multi-view fusion has the most to correct.

Figure 5 shows the same progression qualitatively. Each component contributes a visible improvement: the 2D prior and the cIMLE objective recover finer surface detail, the space-carving loss suppresses the residual cross-view artefacts, and the full model yields the cleanest reconstructions, in agreement with the per-scene metrics above.

The component ablation above is reported on synthetic scenes, where albedo ground truth is available. Table 5 additionally reports the effect of the space-carving loss on real-world novel-view synthesis (Tanks & Temples and Mip-NeRF 360): it improves PSNR, SSIM, and LPIPS on real scenes as well, confirming that the gains are not specific to synthetic data.

2.5 Depth and Novel View Synthesis Comparison with PAPR

To substantiate the PSNR gains of Intrinsic PAPR over the vanilla PAPR [17] reported in the main paper, we compare the two methods on held-out views and

Table 1: Per-scene evaluating novel view synthesis and image quality metrics (PSNR, SSIM, and LPIPS [15]) on synthetic (NeRF Synthetic [11] and TensoIR [5]) datasets, broken down by scene. Higher PSNR and SSIM scores are better, while lower LPIPS scores are better. Our method outperforms the baselines across all metrics on both datasets.

	Chair	Drums	Ficus	Hotdog	Lego	Materials	Mic	Ship	Armadillo	Avg.
PSNR$\uparrow$										
<i>DPIR</i> [3]	26.11	23.16	24.54	25.43	27.35	26.40	27.34	25.50	32.41	26.47
<i>MaterialFusion</i> [10]	30.98	25.26	27.12	27.12	28.62	28.60	30.46	24.77	34.40	28.59
<i>Intrinsic-NeRF</i> [14]	29.37	24.48	28.90	34.58	29.15	26.60	31.13	26.21	35.16	29.51
<i>GS-IR</i> [9]	31.10	24.50	30.77	33.82	32.79	25.79	31.69	26.74	36.40	30.40
<i>IRGS</i> [4]	33.31	24.91	33.83	34.02	32.85	26.46	32.63	27.06	37.11	31.35
<i>DiscretizedSDF</i> [18]	33.42	25.02	34.73	34.36	32.91	27.02	32.89	27.24	37.92	31.72
<i>PAPR</i> [17]	33.59	25.35	36.50	36.40	32.62	29.54	35.64	26.92	38.98	32.84
<i>Intrinsic PAPR (Ours)</i>	37.71	27.81	38.87	36.89	34.95	30.96	36.33	28.16	43.35	35.00
SSIM$\uparrow$										
<i>DPIR</i> [3]	0.822	0.864	0.813	0.805	0.798	0.861	0.946	0.837	0.955	0.867
<i>MaterialFusion</i> [10]	0.902	0.897	0.927	0.936	0.888	0.916	0.958	0.777	0.921	0.902
<i>Intrinsic-NeRF</i> [14]	0.948	0.921	0.958	0.969	0.928	0.935	0.975	0.819	0.972	0.936
<i>GS-IR</i> [9]	0.964	0.928	0.966	0.967	0.963	0.912	0.971	0.857	0.980	0.945
<i>IRGS</i> [4]	0.971	0.932	0.971	0.973	0.969	0.922	0.963	0.862	0.976	0.949
<i>DiscretizedSDF</i> [18]	0.980	0.943	0.982	0.980	0.972	0.934	0.970	0.882	0.990	0.959
<i>PAPR</i> [17]	0.986	0.951	0.994	0.988	0.981	0.972	0.993	0.904	0.989	0.973
<i>Intrinsic PAPR (Ours)</i>	0.990	0.962	0.996	0.991	0.986	0.974	0.995	0.916	0.996	0.978
LPIPS$_{vgg}\downarrow$										
<i>DPIR</i> [3]	0.122	0.145	0.279	0.136	0.140	0.121	0.132	0.113	0.077	0.133
<i>MaterialFusion</i> [10]	0.079	0.105	0.080	0.080	0.111	0.096	0.064	0.262	0.054	0.103
<i>Intrinsic-NeRF</i> [14]	0.061	0.067	0.018	0.031	0.031	0.039	0.025	0.156	0.027	0.051
<i>GS-IR</i> [9]	0.030	0.064	0.028	0.049	0.035	0.079	0.031	0.150	0.032	0.055
<i>IRGS</i> [4]	0.027	0.053	0.039	0.032	0.027	0.049	0.029	0.148	0.029	0.048
<i>DiscretizedSDF</i> [18]	0.025	0.043	0.029	0.029	0.032	0.041	0.025	0.093	0.021	0.038
<i>PAPR</i> [17]	0.018	0.055	0.010	0.021	0.027	0.036	0.007	0.129	0.014	0.035
<i>Intrinsic PAPR (Ours)</i>	0.009	0.033	0.004	0.011	0.013	0.019	0.005	0.082	0.011	0.021

Table 2: Per-scene quantitative comparison of albedo quality on the NeRF Synthetic [11] dataset. For albedo reconstruction, we follow the methods of NeRFactor [16] and GS-IR [9] to compute the scaling factor for each RGB channel. As shown in the table, our albedo reconstruction quality surpasses all baselines across all scenes based on the specified metrics.

	Chair	Drums	Ficus	Hotdog	Lego	Materials	Mic	Ship	Armadillo	Avg.
PSNR$\uparrow$										
<i>DPIR</i> [3]	19.91	20.54	27.02	21.20	20.29	22.51	21.38	20.56	32.62	22.89
<i>Intrinsic-NeRF</i> [14]	18.84	21.97	26.34	25.62	23.00	18.41	30.24	20.90	38.41	24.86
<i>MaterialFusion</i> [10]	22.77	25.69	28.72	27.60	26.58	24.26	26.73	22.60	23.28	25.36
<i>GS-IR</i> [9]	23.91	24.21	30.87	26.75	24.96	22.46	23.31	21.91	38.57	26.33
<i>IRGS</i> [4]	24.50	25.61	31.30	27.83	27.49	26.20	30.30	24.04	38.67	28.44
<i>DiscretizedSDF</i> [18]	25.38	26.12	30.96	27.92	28.13	27.37	30.19	25.14	38.79	28.89
<i>Intrinsic PAPR (Ours)</i>	28.01	29.77	32.08	29.12	30.69	30.74	31.63	29.05	39.80	31.21
SSIM$\uparrow$										
<i>DPIR</i> [3]	0.851	0.796	0.943	0.863	0.849	0.771	0.908	0.706	0.996	0.854
<i>Intrinsic-NeRF</i> [14]	0.855	0.833	0.940	0.962	0.905	0.774	0.959	0.705	0.987	0.880
<i>MaterialFusion</i> [10]	0.862	0.826	0.944	0.968	0.934	0.762	0.949	0.944	0.905	0.899
<i>GS-IR</i> [9]	0.816	0.797	0.948	0.941	0.889	0.850	0.815	0.850	0.986	0.877
<i>IRGS</i> [4]	0.872	0.815	0.947	0.953	0.904	0.862	0.953	0.932	0.981	0.913
<i>DiscretizedSDF</i> [18]	0.868	0.876	0.942	0.960	0.901	0.872	0.962	0.930	0.979	0.921
<i>Intrinsic PAPR (Ours)</i>	0.898	0.931	0.953	0.972	0.935	0.930	0.969	0.960	0.997	0.949
LPIPS$_{Vgg}\downarrow$										
<i>DPIR</i> [3]	0.155	0.182	0.066	0.178	0.849	0.198	0.096	0.198	0.009	0.215
<i>Intrinsic-NeRF</i> [14]	0.137	0.148	0.033	0.097	0.129	0.185	0.051	0.169	0.022	0.108
<i>MaterialFusion</i> [10]	0.130	0.176	0.046	0.050	0.078	0.187	0.048	0.046	0.111	0.097
<i>GS-IR</i> [9]	0.136	0.164	0.053	0.088	0.143	0.208	0.095	0.205	0.051	0.127
<i>IRGS</i> [4]	0.124	0.152	0.042	0.068	0.097	0.152	0.063	0.126	0.034	0.095
<i>DiscretizedSDF</i> [18]	0.130	0.148	0.046	0.072	0.103	0.168	0.075	0.103	0.028	0.097
<i>Intrinsic PAPR (Ours)</i>	0.082	0.115	0.024	0.030	0.068	0.113	0.034	0.032	0.002	0.056

Table 3: Per-scene novel view synthesis and image quality metrics (PSNR, SSIM and LPIPS [15]) on real-world datasets. Columns list all scenes from the Tanks & Temples [7] subset and the real-scene Mip-NeRF 360 dataset [1].

	Tanks & Temples				Mip-NeRF 360								
	Family	Truck	Caterpillar	Ignatius	Barn	Kitchen	Room	Garden	Bicycle	Bonsai	Stump	Counter	Average
PSNR↑													
DPIR [3]	19.86	18.70	19.17	18.18	19.87	22.88	19.91	20.27	18.94	22.83	19.85	21.62	20.17
Intrinsic-NeRF [14]	27.49	23.47	18.80	26.37	23.96	25.69	25.71	23.42	19.73	25.61	21.13	24.11	23.79
MaterialFusion [10]	28.60	24.33	23.06	27.76	23.60	27.10	27.14	23.96	21.86	26.56	22.40	25.75	25.18
GS-IR [9]	30.16	25.84	23.47	26.24	25.52	27.99	28.79	25.72	23.80	28.18	25.37	26.22	26.44
IRGS [4]	33.24	26.21	25.48	26.91	25.93	28.56	29.10	26.54	23.75	28.72	25.78	26.88	27.26
DiscretizedSDF [18]	33.98	27.01	26.28	27.34	26.12	28.72	29.56	27.17	24.10	29.54	26.76	27.73	27.86
PAPR [17]	34.39	26.98	26.79	28.40	27.06	28.31	29.36	28.40	24.21	29.12	26.23	27.13	28.03
Intrinsic PAPR (Ours)	36.38	28.46	27.50	32.18	29.88	29.91	31.19	29.68	25.68	31.35	27.76	29.86	29.99
SSIM↑													
DPIR [3]	0.874	0.844	0.828	0.949	0.791	0.631	0.728	0.693	0.662	0.736	0.646	0.691	0.756
Intrinsic-NeRF [14]	0.920	0.837	0.770	0.917	0.760	0.717	0.781	0.730	0.648	0.722	0.632	0.778	0.768
MaterialFusion [10]	0.974	0.884	0.885	0.880	0.782	0.751	0.792	0.712	0.703	0.756	0.697	0.799	0.801
GS-IR [9]	0.961	0.909	0.878	0.920	0.857	0.867	0.867	0.804	0.706	0.883	0.716	0.839	0.851
IRGS [4]	0.973	0.916	0.912	0.934	0.893	0.889	0.831	0.893	0.756	0.891	0.806	0.847	0.878
DiscretizedSDF [18]	0.979	0.931	0.919	0.942	0.888	0.834	0.812	0.867	0.779	0.894	0.811	0.852	0.876
PAPR [17]	0.983	0.931	0.932	0.956	0.896	0.801	0.822	0.860	0.736	0.807	0.818	0.838	0.865
Intrinsic PAPR (Ours)	0.989	0.943	0.946	0.973	0.922	0.914	0.919	0.905	0.811	0.915	0.831	0.870	0.912
LPIPS _{vgg} ↓													
DPIR [3]	0.154	0.310	0.251	0.101	0.124	0.348	0.404	0.298	0.388	0.311	0.391	0.395	0.290
Intrinsic-NeRF [14]	0.074	0.177	0.218	0.081	0.277	0.306	0.356	0.262	0.342	0.250	0.344	0.335	0.252
MaterialFusion [10]	0.048	0.147	0.137	0.092	0.299	0.268	0.311	0.229	0.299	0.393	0.301	0.321	0.237
GS-IR [9]	0.043	0.142	0.115	0.079	0.179	0.188	0.279	0.158	0.259	0.264	0.258	0.260	0.185
IRGS [4]	0.033	0.112	0.103	0.071	0.160	0.178	0.206	0.152	0.198	0.241	0.199	0.232	0.157
DiscretizedSDF [18]	0.030	0.097	0.093	0.067	0.152	0.163	0.189	0.139	0.181	0.239	0.183	0.231	0.147
PAPR [17]	0.072	0.108	0.157	0.118	0.031	0.180	0.209	0.154	0.201	0.244	0.202	0.226	0.159
Intrinsic PAPR (Ours)	0.018	0.092	0.081	0.045	0.024	0.096	0.112	0.082	0.107	0.171	0.108	0.187	0.094

Table 4: Per-component ablation, per scene. Per-scene novel-view-synthesis PSNR, albedo PSNR, and albedo consistency (MACE) for each rung of the component ablation in the main paper, on the synthetic scenes in our paper. The full model attains the best value on every scene and metric.

	Chair	Drums	Ficus	Hotdog	Lego	Materials	Mic	Ship	Armadillo	Avg.
NVS PSNR$\uparrow$										
Vanilla PAPR	33.59	25.35	36.50	36.40	32.62	29.54	35.64	26.92	38.98	32.84
+ albedo/shading + 2D prior	35.24	26.33	37.45	36.60	33.55	30.11	35.92	27.42	40.73	33.71
+ cIMLE-adapted prior ($M=10$)	36.48	27.07	38.16	36.75	34.25	30.54	36.13	27.79	42.04	34.36
+ space-carving loss	37.45	27.65	38.65	36.86	34.72	30.80	36.27	28.00	43.05	34.83
+ per-view scalar (Full)	37.71	27.81	38.87	36.89	34.95	30.96	36.33	28.16	43.35	35.00
Albedo PSNR$\uparrow$										
+ albedo/shading + 2D prior	26.10	27.60	30.00	27.00	28.60	28.60	29.50	26.70	37.85	29.11
+ cIMLE-adapted prior ($M=10$)	26.67	28.25	30.62	27.64	29.23	29.24	30.14	27.41	38.44	29.74
+ space-carving loss	27.40	29.20	31.45	28.55	30.10	29.95	31.10	28.20	39.30	30.58
+ per-view scalar (Full)	28.01	29.77	32.08	29.12	30.69	30.74	31.63	29.05	39.80	31.21
MACE$\downarrow$										
+ albedo/shading + 2D prior	0.011	0.028	0.029	0.013	0.033	0.015	0.017	0.026	0.004	0.019
+ cIMLE-adapted prior ($M=10$)	0.010	0.026	0.026	0.012	0.030	0.014	0.016	0.024	0.004	0.018
+ space-carving loss	0.009	0.024	0.021	0.010	0.026	0.011	0.015	0.022	0.003	0.016
+ per-view scalar (Full)	0.009	0.023	0.020	0.009	0.025	0.010	0.014	0.021	0.003	0.015

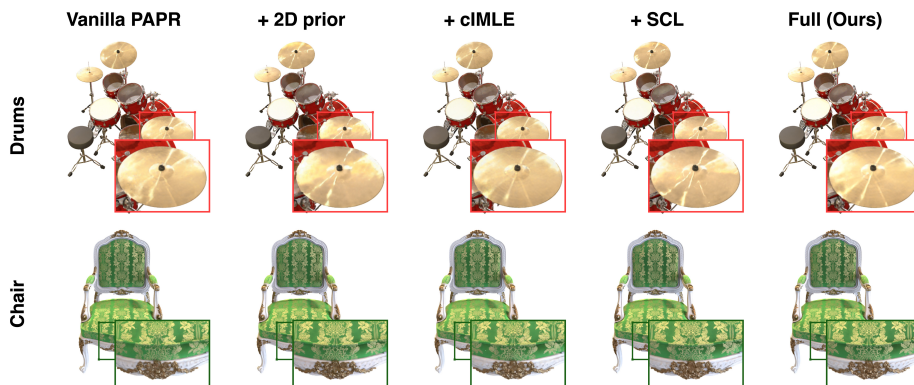


Fig. 5: Qualitative per-component ablation. Novel-view renderings as the components of Intrinsic PAPR are added cumulatively from left to right, starting from vanilla PAPR [17] and ending with our full model, for two representative scenes. Insets magnify a fine-detail region in each scene. Reconstruction quality improves at every stage, consistent with the per-scene numbers in Table 4.

Table 5: Effect of the space-carving loss on real-world novel-view synthesis (Tanks & Temples and Mip-NeRF 360). The space-carving loss improves all three metrics on real scenes as well.

	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$
No SCL	29.14	0.891	0.126
With SCL (Full)	29.99	0.912	0.094

their corresponding depth reconstructions (Figure 6). In the left panels, depth visualizations reveal that PAPR frequently produces blurred, “foggy” geometry with texture-baked structures, while Intrinsic PAPR recovers sharper surfaces and a cleaner scene layout. In the right panels, novel view synthesis results show that these geometric improvements translate into crisper reconstructions with fewer shading and texture artefacts. Taken together, these observations are consistent with our interpretation in Sec. 4 that explicit albedo supervision and intrinsic factorization encourage geometrically faithful reconstructions that benefit both depth and appearance.

2.6 Evaluating the Impact of Pseudo-Ground Truth Albedo Supervision on Baselines

A natural question is whether the improvements of Intrinsic PAPR stem primarily from the additional pseudo-ground-truth albedo supervision provided by the 2D prior, rather than from PAPR’s scene representation itself. To answer this, we conduct a controlled ablation in which every baseline receives the same albedo supervision as our method.

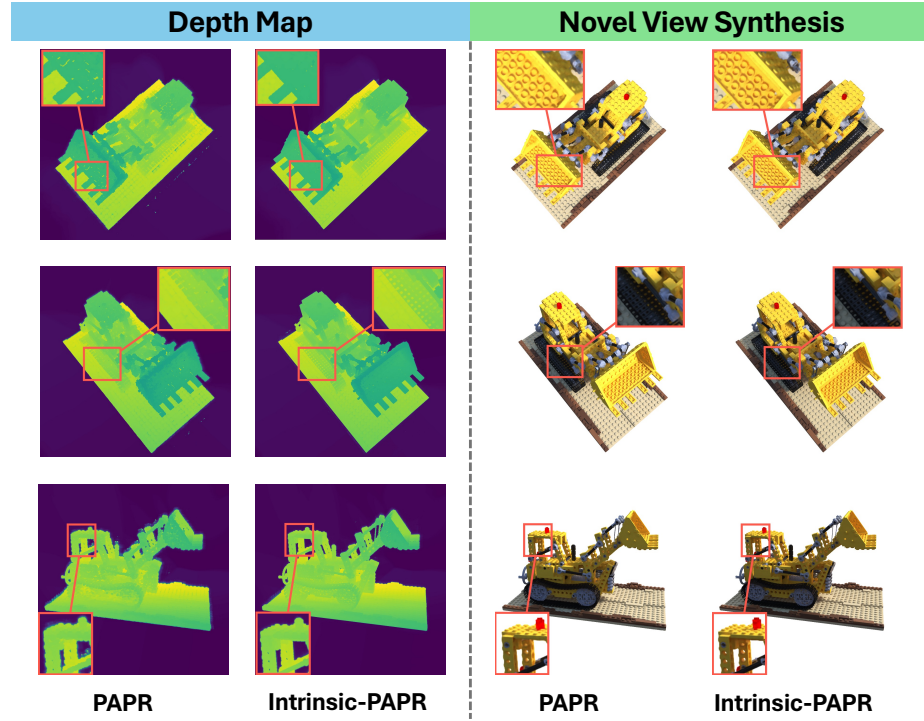


Fig. 6: Qualitative comparison of depth maps and novel view synthesis between PAPER [17] and Intrinsic PAPER. Left: depth reconstructions, where Intrinsic PAPER recovers sharper, less blurry geometry that more closely follows the ground-truth structure than PAPER. Right: corresponding novel views, where this more accurate geometry manifests as clearer image details and reduced shading and texture artefacts.

This experiment is straightforward to set up because all compared baselines (DPIR [3], GS-IR [9], IRGS [4], DiscretizedSDF [18], and Intrinsic-NeRF [14]) are inverse rendering methods that natively predict a diffuse or albedo map as part of their decomposition pipeline. We therefore add an ℓ_2 loss between each baseline’s existing albedo output head and the same per-view pseudo-albedo used by our method, after applying per-view global scale-shift alignment. All other losses, learning rate schedules, and hyper-parameters remain unchanged, ensuring that the only difference from the baselines’ original training is the additional albedo supervision signal.

As shown in Table 6, providing the same pseudo-albedo supervision does not improve reconstruction quality for any of the baselines. This result demonstrates that the additional supervision signal alone is insufficient—performance gains require a representation that can effectively leverage per-view albedo guidance. Specifically, volume-rendering-based methods aggregate features from multiple overlapping primitives along each ray, so per-view pixel-level albedo losses cannot unambiguously supervise individual primitive attributes, and the misattribution issue persists. In contrast, PAPR’s proximity-attention rendering concentrates each pixel’s contribution on a small set of surface points, enabling the albedo supervision to directly constrain per-primitive features and yield the observed improvements.

Table 6: Detailed per-scene quantitative comparison of baselines with and without albedo supervision: Training baseline methods with pseudo-ground truth albedo supervision similar to ours does not improve their reconstruction quality.

	Hotdog			Lego			Materials			Ignatius			Avg.		
	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$
DPIR [3]															
No Albedo Supv.	25.43	0.805	0.136	27.35	0.798	0.140	26.40	0.861	0.121	18.18	0.949	0.101	24.34	0.853	0.124
Albedo Supv.	22.71	0.879	0.163	25.32	0.830	0.166	24.29	0.859	0.150	16.22	0.931	0.143	22.13	0.875	0.155
Intrinsic-NeRF [14]															
No Albedo Supv.	34.58	0.969	0.031	29.15	0.928	0.031	26.60	0.935	0.039	26.37	0.917	0.081	29.18	0.937	0.045
Albedo Supv.	34.06	0.968	0.031	30.76	0.958	0.026	24.82	0.926	0.051	26.24	0.916	0.083	28.97	0.942	0.048
GS-IR [9]															
No Albedo Supv.	33.82	0.967	0.049	32.79	0.963	0.035	25.79	0.912	0.079	26.24	0.920	0.079	29.66	0.941	0.060
Albedo Supv.	26.91	0.950	0.063	29.55	0.946	0.046	25.47	0.919	0.074	26.25	0.920	0.081	27.05	0.934	0.066
IRGS [4]															
No Albedo Supv.	34.02	0.973	0.032	32.85	0.969	0.027	26.46	0.922	0.049	26.91	0.934	0.071	30.06	0.950	0.045
Albedo Supv.	29.23	0.963	0.046	29.54	0.945	0.056	23.87	0.852	0.068	24.56	0.921	0.097	26.80	0.920	0.067
DiscretizedSDF [18]															
No Albedo Supv.	34.36	0.980	0.029	32.91	0.972	0.032	27.02	0.934	0.041	27.34	0.942	0.067	30.41	0.957	0.042
Albedo Supv.	31.23	0.973	0.034	30.02	0.934	0.078	25.02	0.912	0.056	25.78	0.872	0.093	28.01	0.923	0.065

2.7 Albedo Consistency Across Different Views

In Section 5 of the main paper, we introduced the mean albedo consistency error (MACE) to measure how stable the rendered albedo of each surface point remains as the viewpoint changes. Specifically, we track K surface points across T consecutive frames of a rendered camera trajectory, and compute MACE as the

average squared albedo difference between consecutive frames (lower is better). A method with truly view-consistent albedo should produce near-zero MACE, whereas view-dependent shading leakage or misattributed features inflate this error.

Table 7 presents a per-scene breakdown of this metric under three conditions: (i) Intrinsic PAPR *without* the space carving loss (No SCL), (ii) Intrinsic PAPR *with* the space carving loss (With SCL), and (iii) the vanilla 2D prior model [2] used as a reference. Across all scenes, enabling the space carving loss consistently yields the lowest MACE, demonstrating that fusing multi-view albedo hypotheses into a single geometrically consistent representation effectively suppresses view-dependent artefacts. Notably, the variant without space carving produces MACE values comparable to those of the raw 2D prior, confirming that the multi-view consistency enforced by the space carving loss — rather than the albedo supervision alone — is the primary driver of improved cross-view albedo stability.

Table 7: Per-scene Mean Albedo Consistency Error (MACE) for experiments with and without the Space Carving Loss (SCL): Lower MACE values indicate higher albedo consistency. The table shows that, for each scene, using SCL results in lower MACE compared to both the model without SCL and the vanilla 2D prior model [2], indicating enhanced albedo consistency across different views.

Scene	Experiment MACE for Pred. ↓ MACE for vanilla 2D prior		
Lego	No SCL	0.033	0.032
	With SCL	0.025	
Hotdog	No SCL	0.013	0.016
	With SCL	0.009	
Materials	No SCL	0.015	0.015
	With SCL	0.010	
Ficus	No SCL	0.029	0.029
	With SCL	0.020	
Chair	No SCL	0.011	0.013
	With SCL	0.009	
Drums	No SCL	0.028	0.026
	With SCL	0.023	
Mic	No SCL	0.017	0.016
	With SCL	0.014	
Ship	No SCL	0.026	0.025
	With SCL	0.021	
Armadillo	No SCL	0.004	0.004
	With SCL	0.003	

2.8 Decomposition Results on Real-world Scenes

While quantitative albedo evaluation on real-world data is infeasible due to the absence of ground-truth reflectance, we provide qualitative decomposition results to demonstrate that Intrinsic PAPR produces plausible intrinsic factorizations on unconstrained scenes. Figure 7 shows rendered images alongside their corresponding albedo and shading maps for representative scenes from the Mip-NeRF 360 dataset [1]. The decomposed albedo maps successfully remove illumination effects such as cast shadows and specular highlights, recovering spatially uniform reflectance on surfaces with consistent material. The corresponding shading maps capture smooth, geometry-coherent illumination variations without residual texture content, indicating that the factorization does not suffer from the shading leakage commonly observed in volume-rendering-based methods.

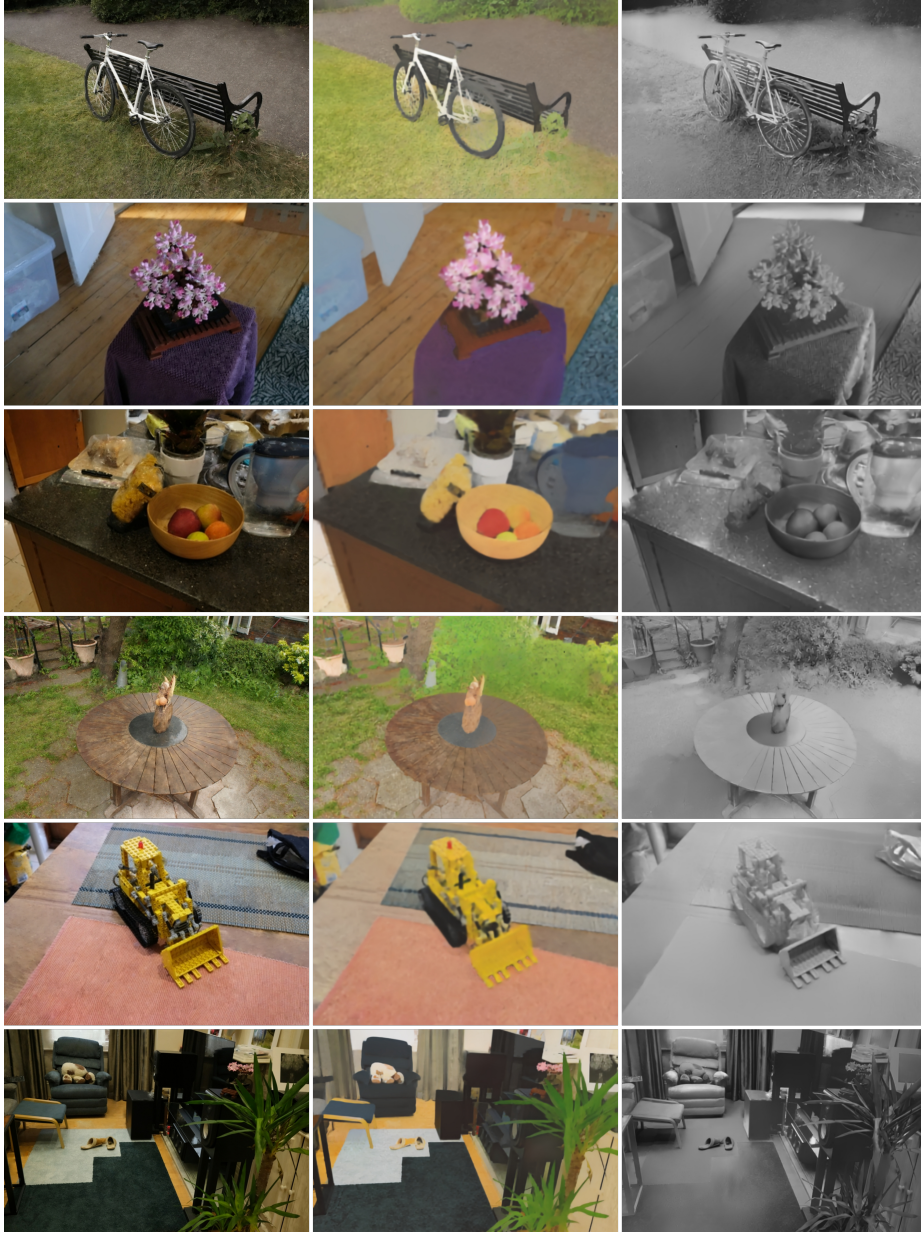


Fig. 7: Qualitative Intrinsic Decomposition on Real-world Scenes. Rendered images alongside their corresponding albedo and shading maps produced by Intrinsic PAPR on representative scenes from the Mip-NeRF 360 dataset [1].

2.9 Albedo Transfer

We present additional qualitative results for the point-level albedo transfer task. Figure 8 showcases transfers on the NeRF Synthetic dataset [11], Figure 10 demonstrates real-world albedo editing on the Tanks & Temples [7] subset, and Figure 9 illustrates the effectiveness of our albedo transfer on unbounded real-world scenes from the Mip-NeRF 360 dataset [1]. Additionally, we include a comparison with Seal-3D [13], a recolouring baseline that directly assigns RGB values through a user-defined brush tool, to demonstrate the broader impact of our intrinsic decomposition-based editing approach. For Seal-3D, we provide a representative qualitative result on the NeRF Synthetic dataset.

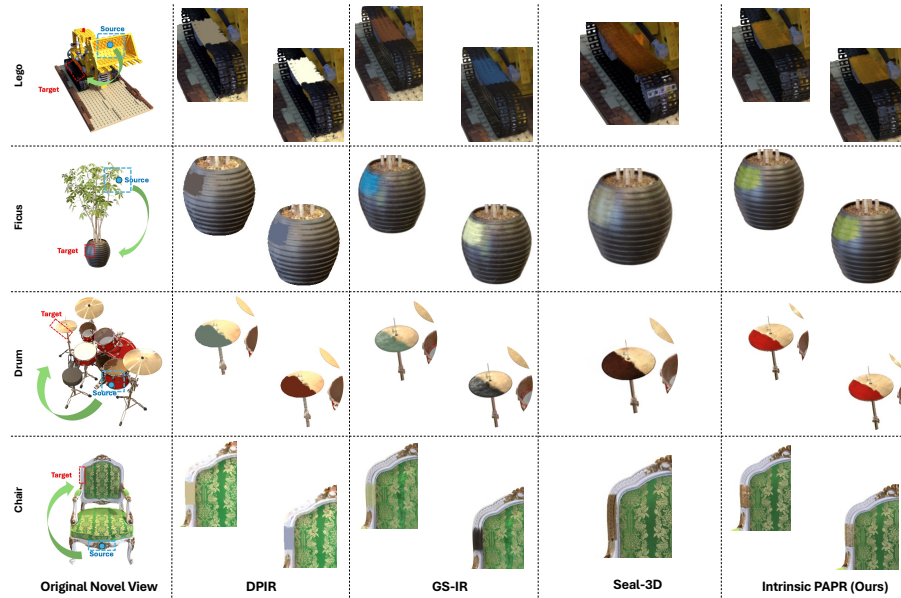


Fig. 8: Additional qualitative results for point-level albedo transfer on NeRF Synthetic [11]. Our approach accurately assigns albedos at the primitive level, reliably transferring them from several randomly chosen source points to the target region. In contrast, baseline methods often produce visually inconsistent outcomes due to the wrong albedo attribution to the primitives.

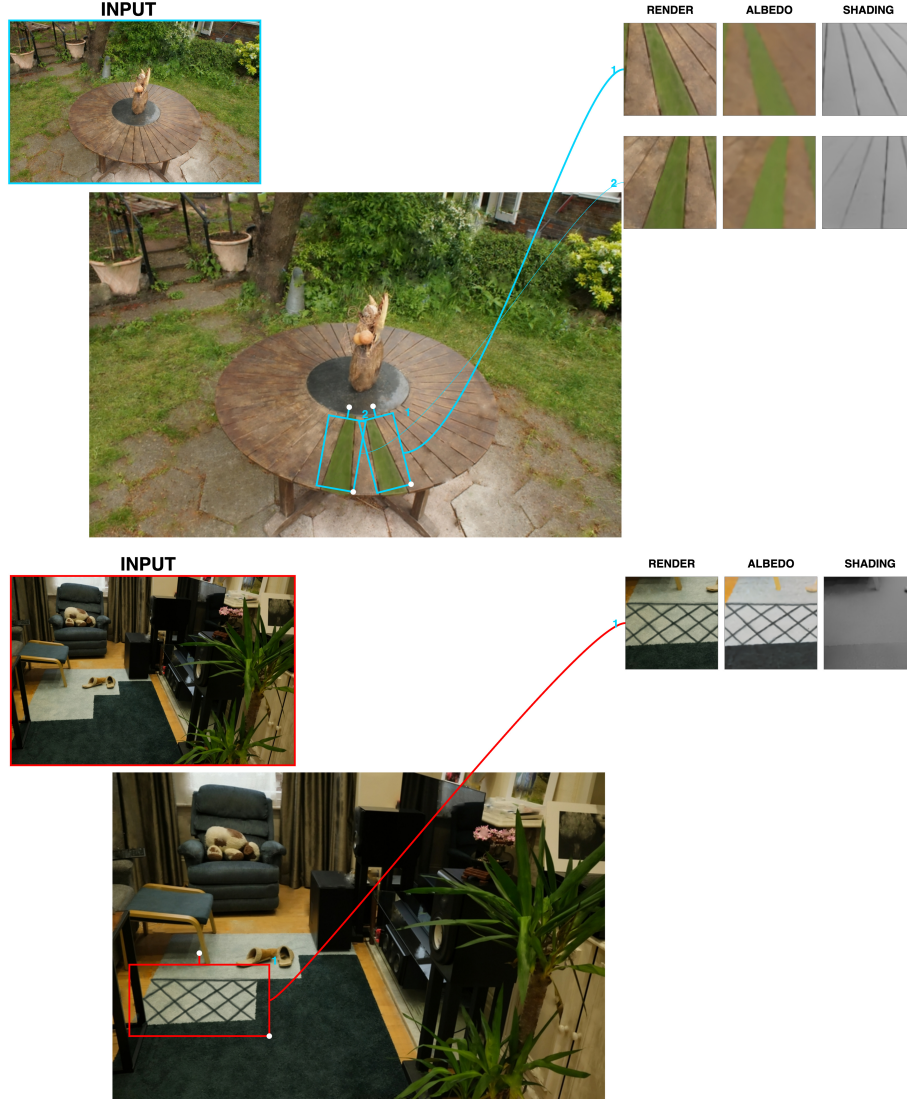


Fig. 9: Albedo Transfer on Mip-NeRF 360 Scenes: Qualitative results for point-level albedo transfer on the Mip-NeRF 360 dataset [1]. In the **Garden** scene (top), the wooden albedo of specific table panels is replaced with a grass colour. In the **Room** scene (bottom), a crisscross pattern is transferred onto a section of a plain rug. In both cases, our method successfully alters the surface albedo while perfectly preserving the original shading and geometry, demonstrating coherent editing in complex, unbounded real-world environments.



Fig. 10: Albedo Transfer on Real Scenes: Comparisons for novel view albedo transfer method on the Tanks & Temples subset [7]. The figure shows results for two scenes (Caterpillar and Family), comparing our method with GS-IR in terms of albedo transfer quality and consistency.

2.10 Shading Transfer

To further showcase our method’s capabilities, we include qualitative results for point-level shading transfer. Figure 11 demonstrates shading edits on the NeRF Synthetic dataset [11], and Figure 12 highlights real-world applicability with shading edits on the Tanks & Temples [7] subset. Furthermore, Figure 13 demonstrates our method’s ability to edit shading in complex, unbounded real-world scenes from the Mip-NeRF 360 dataset [1].

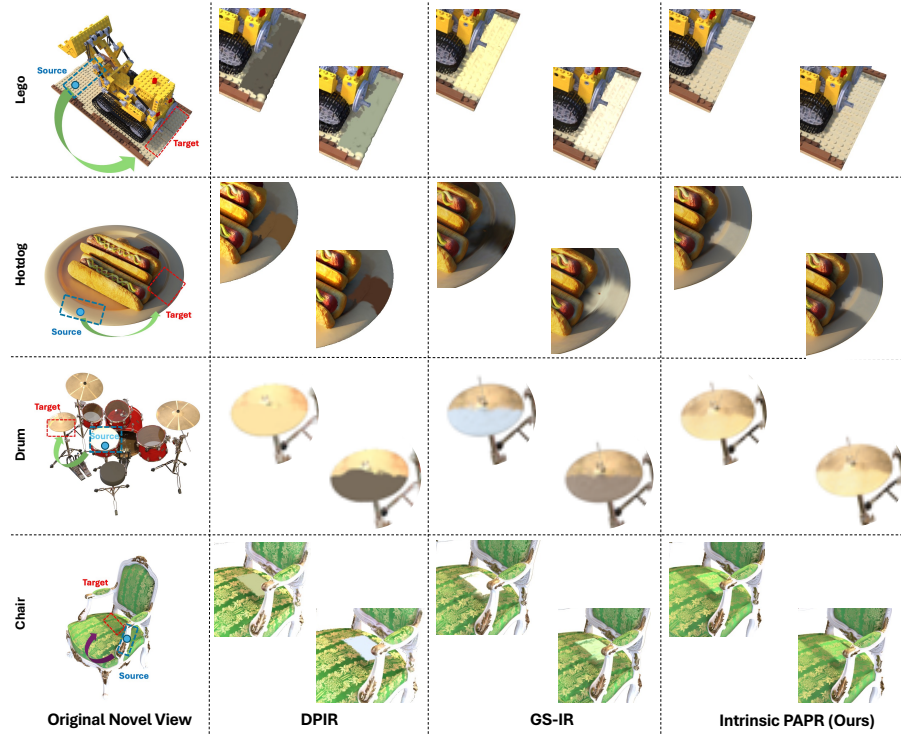


Fig. 11: Additional qualitative comparison of shading transfer results on NeRF Synthetic [11]. Our approach accurately attributes shading at the primitive level, ensuring consistent and precise transfers, while DPIR and GS-IR exhibit inconsistent patterns and visual artefacts due to the wrong shading attribution to the primitives.

2.11 Transfer Error vs. Number of Training Views

To verify that our per-primitive attribution advantage is structural rather than a consequence of the training-data regime, we sweep the number of training views and re-measure the per-primitive transfer error against GS-IR. As Table 8 shows, the advantage is essentially invariant to view count—the albedo-transfer ratio



Fig. 12: Shading Transfer on Real Scenes: Comparisons for novel view shading transfer method on the Tanks & Temples subset [7]. The figure shows results for two scenes (Caterpillar and Family), comparing our method with GS-IR in terms of shading transfer quality and consistency.

Table 8: Per-primitive transfer error (MSE, averaged over the nine synthetic scenes) versus the number of training views. Our advantage over GS-IR is essentially invariant to view count: the albedo-transfer ratio stays within $5.04\text{--}5.23\times$, and at 25 views we already beat GS-IR at 200 views, indicating that the misattribution gap is structural rather than a matter of training data.

		25 views	50 views	100 views	200 views
Albedo	GS-IR	0.398	0.348	0.326	0.314
	Ours	0.078	0.069	0.064	0.060
	Ratio	$5.10\times$	$5.04\times$	$5.09\times$	$5.23\times$
Shading	GS-IR	0.451	0.437	0.424	0.412
	Ours	0.072	0.063	0.058	0.054
	Ratio	$6.26\times$	$6.94\times$	$7.31\times$	$7.63\times$

stays within $5.04\text{--}5.23\times$ across 25 to 200 views, and at 25 views we already beat GS-IR at 200 views—consistent with the main paper: near-coincident primitives render identically from every direction, so additional views cannot disambiguate them.

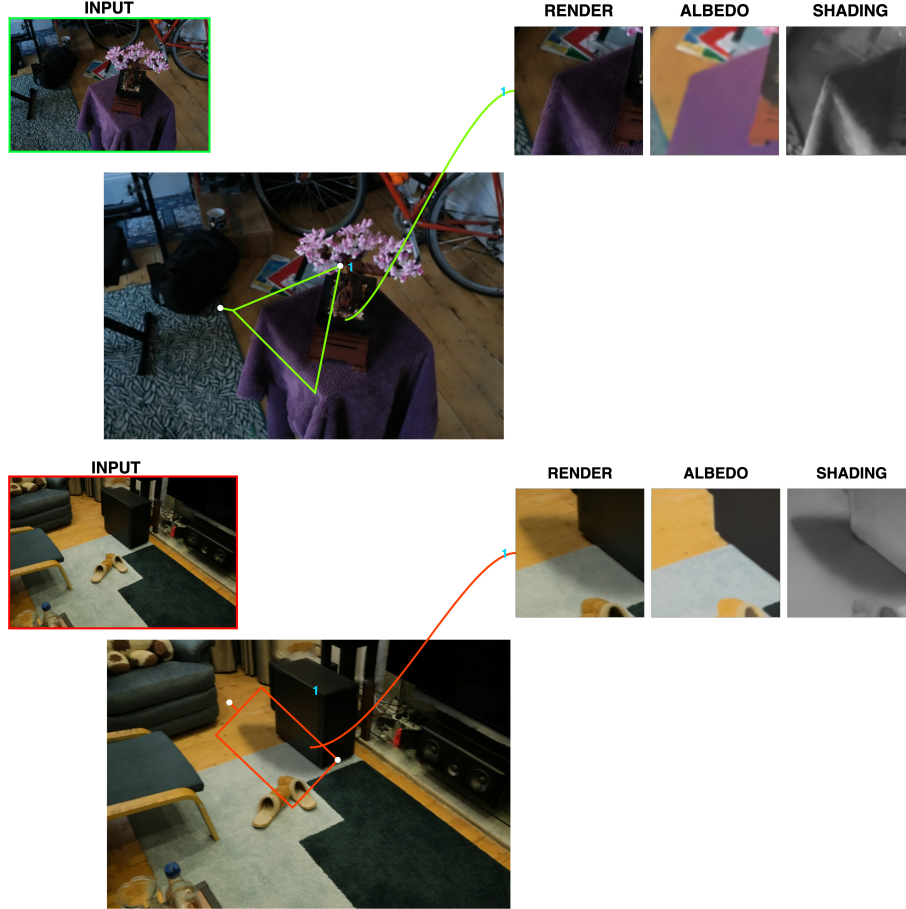


Fig.13: Shading Transfer on Mip-NeRF 360 Scenes: Qualitative results for point-level shading transfer on the Mip-NeRF 360 dataset [1]. In the **Bonsai** scene (top), a distinct triangular shadow is cast onto the purple cloth base. In the **Room** scene (bottom), a synthetic shadow cast by the black speaker is extended over both the wooden floor and the light gray rug. In both cases, our method seamlessly composites the new shadow into the shading map while perfectly maintaining the underlying albedo (reflectance and texture), demonstrating coherent editing in complex, unbounded real-world environments.

2.12 Scene-level Shading Intensity Control

In Figure 14 we present more examples of how our method is able to adjust the overall intensity of shading in a scene.



Fig. 14: Scene-level Shading Intensity Control: This figure showcases our method’s capability to control the overall shading intensity of an image. By building on our accurate shading attribution in the primitive level, we can now modify the shading intensity across the entire image. This is accomplished by scaling the magnitude of shading features for all points in the image by a consistent factor. Decreasing the scaling factor reduces contrast, while increasing it enhances the contrast within the scene.

2.13 Custom Editing Tool

All freeform editing results presented in this paper—including the teaser, the editing figures in Section 4, and the supplementary examples—were produced using a custom graphical user interface that we developed specifically for this work (Figure 15).

The editing workflow proceeds as follows. Given a trained Intrinsic PAPR scene, the user loads a reference view and uses the brush tool to paint over the region they wish to edit (the *target region*). The tool provides fine-grained control over brush colour, size, softness, and shape, as well as an eraser for corrections. The user also selects a *source region* whose albedo or shading attributes should be transferred to the target. Once both regions are defined, the interface exports the corresponding 2D pixel coordinates.

During the subsequent rendering pass, these pixel coordinates are back-projected to identify the underlying 3D surface points in the trained scene representation. The albedo or shading feature vectors of the target points are then replaced with those of the source points—a direct feature transfer that requires no gradient computation, no network re-training, and no further optimization (as detailed in the Computational Cost for Editing section). Because the edit modifies the 3D point features themselves, the changes are automatically multi-view consistent: re-rendering the scene from any novel viewpoint reflects the edit with correct geometry and occlusion handling.

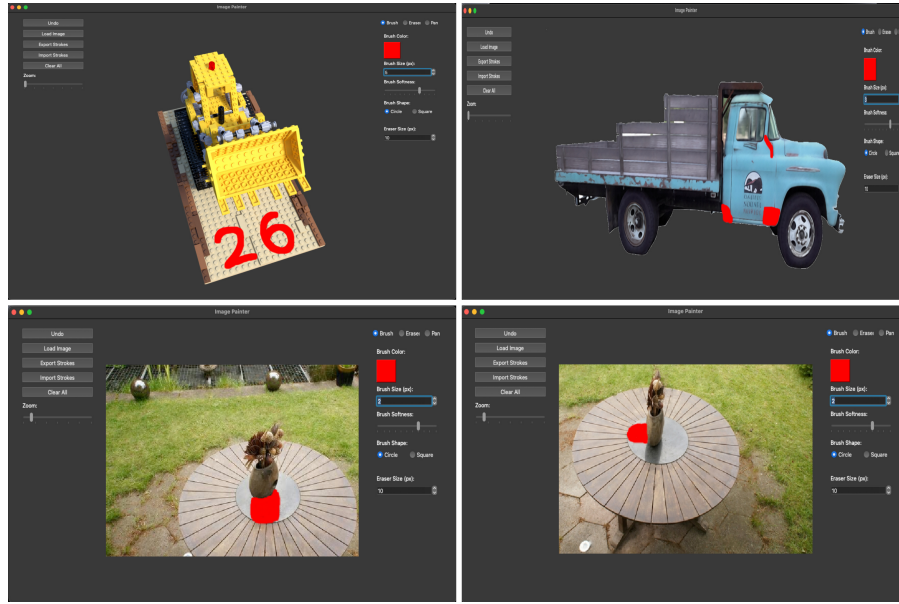


Fig. 15: Custom User Interface for Interactive Scene Editing. This screenshot illustrates our in-house developed tool for freeform intrinsic decomposition editing. Users can load a scene view, then use brush and eraser tools to define target regions. The tool provides controls for brush colour, size, softness, and shape, enabling precise selection of pixels for feature transfer from a specified source region.

2.14 Robustness, Limitations, and Future Work

Out-of-Distribution Robustness. A key practical concern is whether our method generalizes beyond the training distribution of the 2D albedo prior [2], which was trained predominantly on indoor furniture scenes. Our evaluation deliberately covers a broad range of domains: synthetic objects with diverse materials and geometries (NeRF Synthetic [11], TensoIR [5]), large-scale industrial and outdoor scenes (Tanks & Temples [7]), and unbounded real-world environments (Mip-NeRF 360 [1]). Despite significant domain gaps, the method consistently produces high-quality novel view synthesis and plausible albedo decompositions across all benchmarks (Tables 1–3). This robustness is not coincidental: our space carving loss (Sec. 3) explicitly handles prior unreliability by generating multiple albedo hypotheses per view and selecting those that are geometrically consistent across views, effectively filtering out erroneous single-view predictions. Quantitatively, this filtering recovers albedo quality despite the domain gap: relative to the deterministic-prior variant without multi-view filtering, the cIMLE and space-carving filter chain raises albedo PSNR from 29.11 to 31.21 dB and lowers cross-view drift (MACE) from 0.019 to 0.015, confirming that the prior’s unreliable predictions are rejected rather than baked into the 3D albedo.

Robustness on Diverse Materials. Our method deliberately adopts a Lambertian reflectance assumption and a surface-opacity rendering paradigm. To validate the robustness of these design choices under materials that violate them, we evaluate the decomposition on challenging cases encompassing strong specular and complex inter-reflections. Figure 16 presents the predicted albedo and shading for two such scenarios.

Glossy and Specular Surfaces (Drums). The Drums scene contains metallic cymbals (near-mirror) alongside glossy-diffuse drum bodies. On the drum bodies (Figure 16, Regions 3 and 4), the decomposition is clean: specular highlights are correctly attributed to the shading branch, and the albedo recovers a uniform, shadow-free reflectance. On the metallic cymbals (Region 1), the model assigns a low, approximately neutral albedo and absorbs the view-dependent golden reflections into shading. While this is not physically exact (a PBR model would decompose this into metallic and roughness parameters), it yields a *functionally usable* albedo: editing the cymbal’s albedo will not inadvertently alter highlights, because they reside in the shading branch. This case demonstrates that the Lambertian assumption degrades *gracefully* on moderately specular surfaces.

Complex Inter-reflections (Real-world Indoor). The bottom row of Figure 16 combines metallic cookware, glass containers, reflective tabletops, and fabric upholstery in a single real-world frame, exercising multiple assumptions simultaneously. Opaque diffuse surfaces decompose correctly with clean albedo and smooth shading. Specular objects (metal container, Region 1) exhibit graceful degradation similar to the drum scene, with partial environment-reflection leakage into albedo. Additionally, where coloured light bounces between objects (Regions 1 and 4), the ordinal (monochromatic) shading model occasionally attributes the reflected colour to albedo rather than illumination. Despite these

localized artefacts, the overall decomposition remains coherent and does not catastrophically fail, supporting the method’s practical applicability even on complex real-world scenes.

Known Limitations on Extreme Materials. While our model exhibits robust practical performance, predominantly specular surfaces (e.g., perfect mirrors) and highly transparent objects (e.g., clear glass, as seen in Region 3 of the indoor scene) represent the expected boundaries of our assumptions. For near-mirror surfaces, the space carving loss cannot perfectly isolate the minimal diffuse reflectance, leading to environment reflections partially leaking into the albedo. For transparent objects, PAPR’s surface-aligned proximity attention lacks a mechanism for compositing content visible *through* a surface, causing the albedo to absorb a blend of the foreground tint and the background. As discussed in the main text, these are necessary architectural trade-offs accepted in exchange for resolving the misattribution issue on the vast majority of opaque, primarily diffuse real-world surfaces.

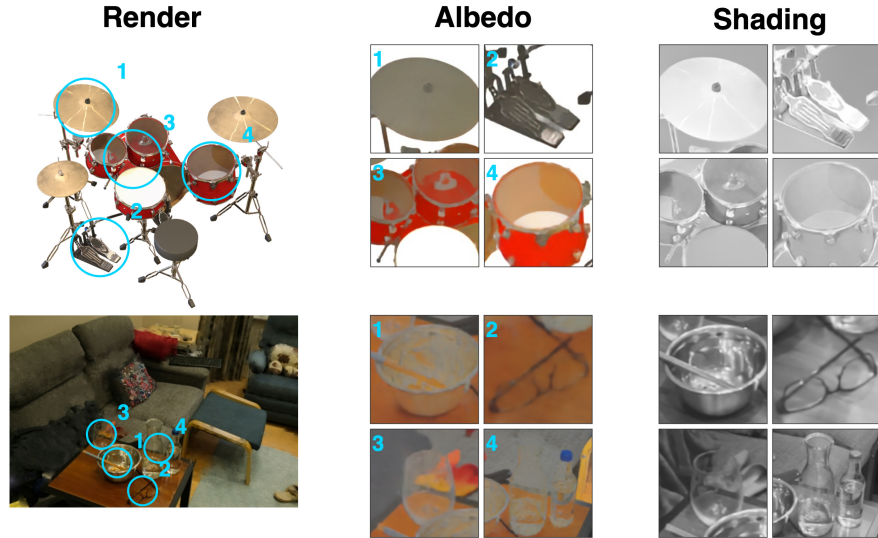


Fig. 16: Decomposition on Challenging Materials. We evaluate our method on scenes that stretch the limits of our Lambertian and surface-opacity assumptions. **Top (Drums):** On glossy-diffuse drum bodies (Regions 3, 4), the decomposition successfully recovers uniform albedo while pushing specular highlights into shading. On near-mirror cymbals (Region 1), the method functionally absorbs the golden reflections into shading, leaving a neutral albedo. **Bottom (Indoor):** In a complex real-world setting, opaque diffuse surfaces decompose cleanly. Localized artefacts appear at glass boundaries (Region 3) due to the surface-opacity assumption, and coloured inter-reflections (Region 1, 4) occasionally bleed into albedo due to the ordinal shading prior. However, the overall scene composition remains highly coherent.

Ordinal Shading and Non-Monochromatic Illumination. The 2D prior model [2] supports both ordinal (greyscale) and colourful (chromatic) shading decomposition. In the current implementation, we use its *ordinal shading* output, which assumes monochromatic illumination. This is a deliberate design choice that simplifies the albedo-shading factorization and is appropriate for the majority of scenes under neutral or white lighting.

However, under strongly coloured illumination, this assumption causes the light colour to be absorbed into the predicted albedo rather than the shading component. Figure 17 illustrates this effect: when a green point light is added to a synthetic scene, both the raw 2D prior and our method bake the green tint into the albedo. Importantly, this is not a fundamental limitation of the Careaga et al. prior—it is a consequence of our current choice to use the ordinal output. Upgrading to the colourful shading branch of the same prior model would directly resolve this artefact, and we plan this as an immediate extension.

Future Directions. The most immediate extension is replacing the ordinal shading output with the colourful shading branch of the same 2D prior [2], which would enable decomposition under non-monochromatic illumination with minimal architectural changes. Beyond this, our prior-based supervision paradigm naturally accommodates richer material properties: additional per-point feature channels for roughness or metallicity can be introduced alongside PBR-aware priors, following the same supervision strategy used for albedo. More broadly, as 2D intrinsic decomposition models continue to advance, our framework can directly benefit from more accurate and domain-general priors without requiring changes to the 3D representation or training procedure.

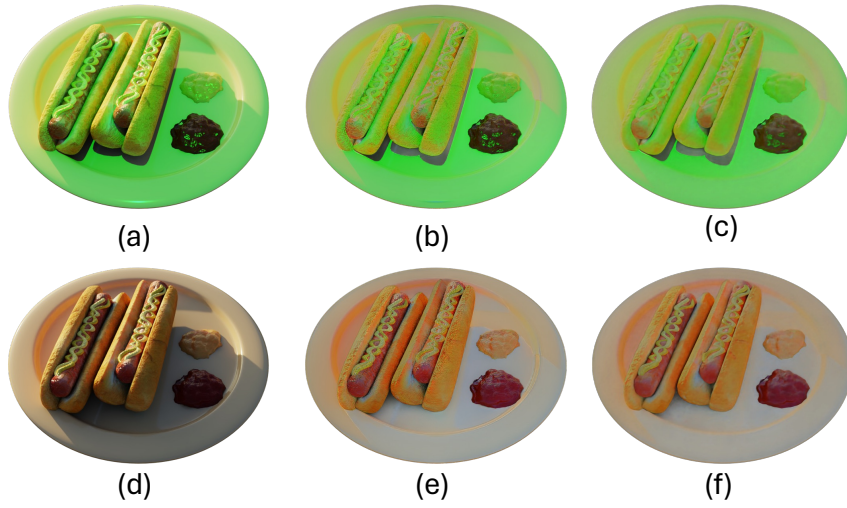


Fig.17: Comparison of Albedo Extraction Under Modified and Original Lighting Conditions. (a) shows the input render with an added green point light. In (b), the vanilla intrinsic decomposition model using ordinal shading baked the green light into the albedo. (c) displays our model’s predicted albedo, which inherently baked the light colour to albedo. Under the original lighting setup, panel (d) presents the input view, (e) shows the albedo extracted by the same 2D prior model, and (f) illustrates our model’s predicted albedo. This figure highlights how biases in the pseudo-ground-truth data can propagate to our training.

References

1. Barron, J.T., Mildenhall, B., Verbin, D., Srinivasan, P.P., Hedman, P.: Mip-NeRF 360: Unbounded anti-aliased neural radiance fields. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5460–5469. IEEE (2022)
2. Careaga, C., Aksoy, Y.: Colorful diffuse intrinsic image decomposition in the wild. *ACM Transactions on Graphics* **43**(6) (2024)
3. Chung, H.G., Choi, S., Baek, S.H.: Differentiable point-based inverse rendering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4399–4408 (2024). <https://doi.org/10.1109/CVPR52733.2024.00421>
4. Gu, C., Wei, X., Zeng, Z., Yao, Y., Zhang, L.: IRGS: Inter-reflective gaussian splatting with 2D gaussian ray tracing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10943–10952 (2025). <https://doi.org/10.1109/CVPR52734.2025.01022>
5. Jin, H., Liu, I., Xu, P., Zhang, X., Han, S., Bi, S., Zhou, X., Xu, Z., Su, H.: TensoIR: Tensorial inverse rendering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 165–174 (2023). <https://doi.org/10.1109/CVPR52729.2023.00024>
6. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: Bengio, Y., LeCun, Y. (eds.) 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7–9, 2015, Conference Track Proceedings (2015), <https://arxiv.org/abs/1412.6980>, accessed 2026-07-01
7. Knapitsch, A., Park, J., Zhou, Q.Y., Koltun, V.: Tanks and temples: Benchmarking large-scale scene reconstruction. *ACM Transactions on Graphics* **36**(4), 78:1–78:13 (2017)
8. Li, K., Peng, S., Zhang, T., Malik, J.: Multimodal image synthesis with conditional implicit maximum likelihood estimation. *International Journal of Computer Vision* **128**(10–11), 2607–2628 (2020)
9. Liang, Z., Zhang, Q., Feng, Y., Shan, Y., Jia, K.: GS-IR: 3D gaussian splatting for inverse rendering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 21644–21653 (2024). <https://doi.org/10.1109/CVPR52733.2024.02045>
10. Litman, Y., Patashnik, O., Deng, K., Agrawal, A., Zawar, R., la Torre, F.D., Tulsiani, S.: MaterialFusion: Enhancing inverse rendering with material diffusion priors. In: 2025 International Conference on 3D Vision (3DV). pp. 802–812. IEEE (2025). <https://doi.org/10.1109/3DV66043.2025.00079>
11. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: NeRF: Representing scenes as neural radiance fields for view synthesis. In: European Conference on Computer Vision (ECCV). Springer, Cham, Switzerland (2020)
12. Peng, S., Moazenipourasil, S.A., Li, K.: CHIMLE: Conditional hierarchical IMLE for multimodal conditional image synthesis. *Advances in Neural Information Processing Systems* **35**, 280–296 (2022)
13. Wang, X., Zhu, J., Ye, Q., Huo, Y., Ran, Y., Zhong, Z., Chen, J.: Seal-3D: Interactive pixel-level editing for neural radiance fields. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 17637–17647 (2023). <https://doi.org/10.1109/ICCV51070.2023.01621>

14. Ye, W., Chen, S., Bao, C., Bao, H., Pollefeys, M., Cui, Z., Zhang, G.: Intrinsic-NeRF: Learning intrinsic neural radiance fields for editable novel view synthesis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 339–351 (2023). <https://doi.org/10.1109/ICCV51070.2023.00038>
15. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18–22, 2018. pp. 586–595. Computer Vision Foundation / IEEE Computer Society (2018). <https://doi.org/10.1109/CVPR.2018.00068>
16. Zhang, X., Srinivasan, P.P., Deng, B., Debevec, P., Freeman, W.T., Barron, J.T.: NeRFactor: Neural factorization of shape and reflectance under an unknown illumination. *ACM Transactions on Graphics (ToG)* **40**(6), 1–18 (2021)
17. Zhang, Y., Peng, S., Moazeni, A., Li, K.: PAPR: Proximity attention point rendering. In: Advances in Neural Information Processing Systems 36 (NeurIPS 2023) (2023), https://proceedings.neurips.cc/paper_files/paper/2023/hash/bda5c35eded86adaf0231748e3ce071c-Abstract-Conference.html, accessed 2026-07-01
18. Zhu, Z.L., Yang, J., Wang, B.: Gaussian splatting with discretized SDF for re-lightable assets. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 25155–25164 (2025). <https://doi.org/10.1109/ICCV51701.2025.02333>