INVSPLAT: INVERSE FEED-FORWARD SCENE SPLATTING

Anonymous authors
Paper under double-blind review

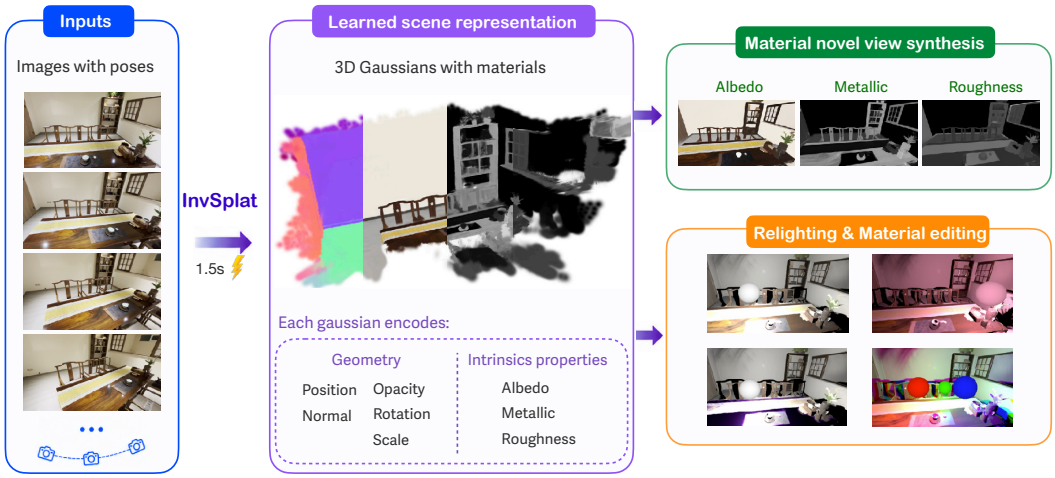

Figure 1: InvSplat Overview. Given a set of posed images, InvSplat reconstructs both the 3D scene geometry and material parameters within seconds, enabling novel view synthesis and relighting.

ABSTRACT

Inverse rendering recovers geometry and material properties from images and is a prerequisite for relighting and material editing. Recent learning-based methods predict materials from images, but despite exploiting multi-view information, their image-space predictions can remain inconsistent across viewpoints. Meanwhile, per-scene optimization methods are computationally expensive and, in our experiments, achieve lower material accuracy than learned approaches despite using many views and can bake illumination into material maps. We therefore predict materials directly in 3D, as part of a feed-forward Gaussian splatting reconstruction. We observe that a direct combination of a 2D material estimator with a feed-forward reconstruction model is insufficient: the two networks must share features, and surface normals derived from the reconstructed geometry absorb texture, which degrades relighting. Based on these observations, we present InvSplat, which reconstructs 3D Gaussians with albedo, metallic, roughness and normals from posed images in 1.5 seconds. InvSplat predicts normals from appearance features and uses a point-light relighting loss that jointly constrains materials and normals without introducing errors from reconstructed geometry into the shading supervision. From two input views, InvSplat achieves higher relighting PSNR than the evaluated 2D baselines and produces more temporally consistent materials than methods processing all 32 frames of a video sequence, while maintaining comparable per-view material accuracy. Its explicit 3D representation further enables novel-view rendering of materials, relighting, and material editing.

1 INTRODUCTION

Inverse rendering estimates scene geometry and material properties from images, enabling relighting and material editing. Existing approaches broadly fall into two categories. Optimization-based methods (Liang et al., 2024; Lin et al., 2025; Zhang et al., 2021; Azinovic et al., 2019; Boss et al.,

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

2021) fit a 3D scene representation to each capture, typically requiring many views and 30 minutes to two hours per scene in our experiments. Learning-based methods (Li & Snavely, 2018; Li et al., 2020; Zhu et al., 2022; Wu et al., 2025; Liang et al., 2025; Zheng et al., 2025) instead predict material maps in a forward pass. Although recent models process multiple views jointly, they produce image-space estimates without an explicit shared scene representation. As a result, material predictions for the same surface can differ across views, causing flickering in video sequences, particularly in regions with specular highlights (Tab. 2 and Fig. 3). We find that fine-tuning MVInverse on the same training data does not resolve this inconsistency in metallic and roughness (Tabs. 1, 2 and 6). Conversely, the evaluated optimization-based methods recover less accurate materials despite fitting a shared representation to multiple views (Tab. 5).

These observations motivate combining learned material estimation with a shared 3D representation. Feed-forward 3D reconstruction methods (Charatan et al., 2024; Chen et al., 2024; Xu et al., 2025a) provide such a representation: they predict 3D Gaussians from a few posed images in a single forward pass. However, they typically model appearance using RGB or spherical harmonics, which bake in illumination and do not directly support relighting (Fig. 7). A natural approach is therefore to combine a feed-forward reconstruction model with a 2D material estimator and to assign the predicted materials to the Gaussians. We find that this approach has two shortcomings. First, running the two pretrained networks side by side performs worse than a model in which geometry and material branches exchange features (Tab. 3). Second, surface normals obtained from the reconstructed geometry, contain texture from the input images, which leads to incorrect shading under new illumination (App. E). In addition, supervising albedo, metallic, roughness and normals separately does not ensure that their combination produces correct shading.

We address these problems with InvSplat, a feed-forward model that reconstructs a 3D Gaussian scene with physically based (PBR) materials from posed images. The geometry and material branches are initialized from pretrained networks and connected so that material features inform depth estimation and image features inform albedo prediction. A dedicated head predicts shading normals from appearance features rather than deriving them from reconstructed geometry. To jointly constrain materials and normals, we apply point-light relighting supervision to Gaussian-rendered intrinsics. Both predicted and ground-truth intrinsics are shaded using ground-truth geometry, isolating the loss from reconstructed-depth errors.

InvSplat reconstructs a scene in 1.5 seconds and matches the per-view accuracy of state-of-the-art 2D methods while producing more consistent materials across views. On video sequences, InvSplat reconstructs the scene from the first and last frame only and still achieves lower temporal error than 2D methods that process all 32 frames, reducing albedo error by a factor of 1.3 to 3.9. Compared to per-scene optimization, it is more accurate and more than three orders of magnitude faster under sparse view settings. The explicit representation further supports rendering materials from novel views, relighting with cast shadows and material editing.

We summarize our contributions as follows:

- We extend feed-forward 3D Gaussian reconstruction to material estimation from as few as two input views. By representing materials in a shared 3D scene rather than predicting separate image-space maps, our method improves cross-view consistency and enables novel-view material rendering compared with 2D methods that use multi-view conditioning.
- Our experiments demonstrate that coupling geometry and intrinsic features improves material estimation over directly combining pretrained reconstruction and material-estimation networks, and that predicting normals from appearance features avoids texture artifacts observed in geometry-derived normals.
- We further demonstrate that applying point-light relighting supervision to Gaussian-rendered materials and normals improves relighting quality.

2 Related work

Optimization/Learning-based approaches. Optimization-based inverse rendering fits a scene representation to each capture via iterative updates. It can be accurate and physically grounded, but inference is slow and must be repeated for every new scene. For example, IRIS (Lin et al., 2025) estimates geometry, physically based materials, spatially varying HDR lighting, and camera

response from posed multi-view LDR images via an iterative optimization pipeline. Intrinsic Image Fusion (Kocsis et al., 2026) is another optimization-based multi-view approach that fuses per-view intrinsic priors into a consistent, low-dimensional material space and then refines it with inverse path tracing. To improve material estimation, learning-based methods also supervise the appearance of predicted intrinsics under novel lighting: Prakash et al. (2022) use ground-truth geometry and normals, while HumanMaterial (Jiang et al., 2025b) also refines predicted normals in image space. We apply relighting supervision to Gaussian-rendered materials and normals, allowing gradients to reach geometry through rasterization and linking training to our inference-time relighting model.

Generative approaches. Diffusion models provide strong image priors for inverse problems and have been used to predict intrinsic properties or G-buffers. DiffusionRenderer (Liang et al., 2025) leverages video diffusion priors to estimate G-buffers from real videos and additionally trains a neural renderer from G-buffers to enable editing workflows; however, it inherits the computational cost and stability challenges of diffusion sampling. DNF-Intrinsic (Zheng et al., 2025) proposes a deterministic alternative (flow-matching / noise-free diffusion) together with a generative rendering constraint to improve faithfulness, but it remains substantially heavier than a single forward pass. Moreover, without explicit 3D coupling, such image-space diffusion approaches can produce predictions that drift across viewpoints. GenRecon (Schmid et al., 2026) casts multi-view scene reconstruction as conditional 3D generation with a pretrained shape generator, producing editable PBR meshes. Its outputs are constrained to the generator's latent space and require iterative sampling per scene chunk, whereas InvSplat predicts materials directly from the observed pixels in a single forward pass.

Feed-forward scene reconstruction. Recent advancements in 3D scene representation have seen a paradigm shift towards feed-forward 3D Gaussian Splatting (3DGS) architectures, which bypass time-consuming per-scene optimization (Kerbl et al., 2023) to generate high-fidelity novel views. pixelSplat (Charatan et al., 2024) pioneered this direction by learning to reconstruct radiance fields from sparse image pairs using dense probability distributions. Subsequent works rapidly improved both efficiency and geometric accuracy; MVSplat (Chen et al., 2024) introduced plane-sweeping cost volumes for fast multi-view feature matching, while DepthSplat (Xu et al., 2025b) integrated pre-trained depth features to inject stronger geometric priors. To address the limitations of single-pass inference, ReSplat (Xu et al., 2025a) proposed a recurrent framework that iteratively refines 3D Gaussian primitives using rendering error feedback. More recently, researchers have expanded input constraints and reconstruction objectives, with Anysplat (Jiang et al., 2025a) directly predicting scene Gaussians from unposed images and WorldMirror (Liu et al., 2025) jointly predicting multiple 3D representations. These methods target RGB novel-view synthesis and do not model intrinsic scene properties. In contrast, InvSplat predicts geometry and PBR materials in a single forward pass, in a 3D representation shared across views.

3 METHOD

Given N input images $\{I_i\}_{i=1}^N$ capturing a shared scene and their corresponding camera parameters $\{\mathbf{P}_i\}_{i=1}^N$ (including intrinsics and extrinsics), our goal is to jointly recover the scene geometry $\mathcal{G}$ and intrinsic material properties $\mathcal{M}$ in a view-consistent 3D representation that supports physically based relighting. We employ a feed-forward network f_θ that predicts all scene parameters in a single forward pass:

$$(\mathcal{G}, \mathcal{M}) = f_\theta\big(\{I_i\}_{i=1}^N, \{\mathbf{P}_i\}_{i=1}^N\big), \tag{1}$$

Sec. 3.1 introduces our material-aware Gaussian scene representation. Sec. 3.2 describes the the dual-branc architecture that predicts it. Sec. 3.3 shows how the predicted scene can be relit under novel illumination, and Sec. 3.4 presents the training objectives, including a relighting loss built on the same shading model.

3.1 SCENE REPRESENTATION WITH INTRINSIC PROPERTIES

We represent the scene with M 3D Gaussian primitives (Kerbl et al., 2023). Each Gaussian j is parameterized by its position $\boldsymbol{\mu}_j \in \mathbb{R}^3$, rotation quaternion $\mathbf{q}_j \in \mathbb{R}^4$, scale $\mathbf{s}_j \in \mathbb{R}^3$, and opacity $\sigma_j \in [0, 1]$. The original formulation models appearance with spherical harmonics coefficients, which entangle material properties with captured illumination. We instead associate each Gaussian with light-independent intrinsic material properties: albedo $\mathbf{a}_j \in [0, 1]^3$, metallic $m_j \in [0, 1]$, and

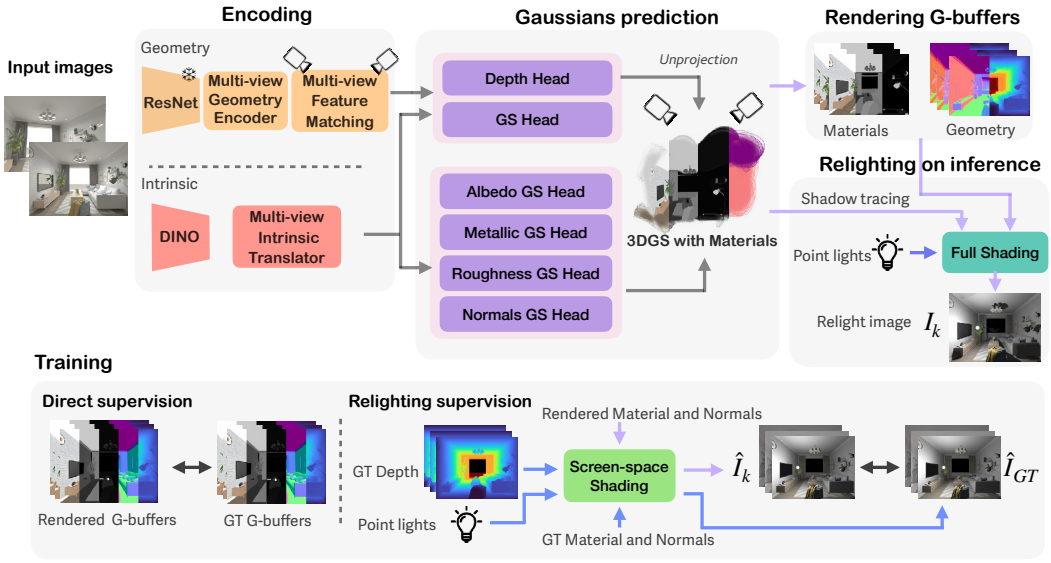

Figure 2: Given posed multi-view images, InvSplat predicts in a single forward pass a 3D Gaussian scene representation with material parameters (albedo, metallic, roughness). The predicted scene is supervised through rendered G-buffers and point-light relighting on ground-truth depth. At inference, we relight the reconstructed scene under arbitrary point-light with full-shading, using ray tracing to cast shadows.

roughness $r_j \in [0, 1]$. We further augment each Gaussian with an explicit unit surface normal $\mathbf{n}_j \in \mathbb{R}^3$, which is required for physically based rendering. The scene is thus represented as

$$\mathcal{G} = \{(\boldsymbol{\mu}_j, \mathbf{q}_j, \mathbf{s}_j, \sigma_j, \mathbf{n}_j)\}_{j=1}^{M}, \qquad \mathcal{M} = \{(\mathbf{a}_j, m_j, r_j)\}_{j=1}^{M}.$$

3.2 Feed-forward inverse rendering architecture

Geometry and materials depend on complementary cues. Accurate geometry is best recovered from explicit multi-view correspondence. Material estimation, by contrast, requires high-level priors that resolve ambiguities that correspondence alone cannot, such as whether a dark region is a dark material or a bright material in shadow. Our dual-branch design combines these complementary sources of information (Fig. 2). We build on ReSplat (Xu et al., 2025a) and retain its pretrained multi-view matching branch as our *geometry branch*, so that predicted Gaussians inherit its view-consistent structure. Its monocular branch is replaced by an *intrinsic branch* derived from MVInverse (Wu et al., 2025), which predicts materials from multi-view images without an explicit 3D representation and thus contributes a strong material prior that our model grounds in 3D. The two branches share features during decoding: intrinsic features contribute to depth estimation, while the high-resolution image features of the geometry branch sharpen albedo prediction. All per-view predictions are finally unprojected with the predicted depth into a set of 3D Gaussians with intrinsic material attributes.

Geometry branch. A ResNet (He et al., 2015) backbone maps each view I_i to a multi-scale feature pyramid $\{\mathbf{F}_i^s\}$. The deepest pyramid level is processed by the *Multi-view Geometry Encoder*, a transformer that attends across views to produce view-aware geometry features. The *Multi-view Feature Matching* module then warps these features between views using the known cameras $\{\mathbf{P}_i\}$ and builds a depth-candidate cost volume $\mathbf{C}_i$ that encodes cross-view correspondence. The shallower pyramid levels bypass the matching step and are forwarded directly to the decoding heads, providing the high-resolution image cues needed for dense per-pixel prediction.

Intrinsic branch. Initialized from MVInverse (Wu et al., 2025), this branch extracts appearance features using a DINOv2 (Oquab et al., 2023) ViT-L/14. Each view is encoded independently into patch features, which are refined by the *Multi-view Intrinsic Translator*, a 36-block transformer that alternates intra-view and inter-view self-attention (Wang et al., 2025a). In contrast to the geometry branch, views exchange information only through attention, never through explicit warping, leaving

the branch free to draw on appearance priors rather than precise correspondence. Features $\{\mathbf{F}_i^m\}$ from uniformly spaced translator layers feed both the material heads and the depth head, so they are trained jointly by both and serve as a representation shared between materials and geometry.

Decoding and Gaussian construction. Following ReSplat, a UNet predicts depth d_i over depth candidates from $\mathbf{C}_i$, $\{\mathbf{F}_i^m\}$ and $\{\mathbf{F}_i^s\}$, upsampled by a DPT module (Ranftl et al., 2020). We lift features into 3D using predicted depth and refine them with a Point Transformer (Zhao et al., 2021; Xu et al., 2025a), followed by a GS head that predicts Gaussian rotation, scale, and opacity. Four DPT heads initialized from MVInverse predict albedo, metallic, roughness, and surface normals from intrinsic features. For albedo prediction, we replace MVInverse's separate ResNeXt features with high-resolution features from our geometry branch, eliminating the redundant CNN backbone. Finally, we unproject the predicted depth maps to obtain Gaussian centers $\{\boldsymbol{\mu}_j\}_{j=1}^M$. Each Gaussian inherits the predicted material properties $(\mathbf{a}_j, m_j, r_j)$ from its source pixel, while its camera-space normal is transformed into world coordinates using the input camera extrinsics.

Rendering. We use a differentiable Gaussian splatting renderer to project the predicted scene into arbitrary viewpoints, producing albedo, metallic, roughness, normal, and depth maps (G-buffers) in a single rasterization pass.

3.3 Inference-time Relighting

Our explicit material representation enables physically based relighting of reconstructed indoor scenes. We model indoor lighting using point lights, allowing us to model spatially varying shading and cast shadows from sources such as lamps. Using the rendered G-buffers, we recover per-pixel surface positions $\mathbf{x}_p$ from depth and trace visibility $V_k(p)$ toward each light using BVH-accelerated ray tracing through the Gaussians, following (Gao et al., 2024). The outgoing direct radiance is

$$\mathbf{L}_o(p) = \sum_{k=1}^{K} V_k(p)\, f_{\mathrm{CT}}(\boldsymbol{\omega}_o, \boldsymbol{\omega}_k; \mathbf{a}_p, m_p, r_p, \mathbf{n}_p) \odot \frac{\Phi_k \mathbf{c}_k}{\|\mathbf{l}_k - \mathbf{x}_p\|_2^2 + \rho^2/2} \max\big(0, \mathbf{n}_p^\top \boldsymbol{\omega}_k\big)\,, \tag{2}$$

where K is the number of point lights, and f_{CT} is the Cook–Torrance BRDF (Cook & Torrance, 1982) with Lambertian diffuse and GGX specular components, evaluated using the rendered materials and normals. Light k has position $\mathbf{l}_k$, intensity Φ_k, and color $\mathbf{c}_k$; $\boldsymbol{\omega}_k$ and $\boldsymbol{\omega}_o$ point toward the light and camera. The parameter ρ regularizes attenuation, and $\odot$ denotes channel-wise multiplication. Per-pixel visibility avoids blurring shadow boundaries across Gaussians. We reuse the BVH as lights or cameras move and convert the shaded image to sRGB.

3.4 Training

During training, for each forward path, given N_t camera poses, we render N_t maps for each geometric and material property and supervise all the rendered properties and intermediate predicted depth. For each view i, let $(\mathbf{a}_i, m_i, r_i, d_i, \mathbf{n}_i)$ denote the ground-truth albedo, metallic, roughness, depth, and normals, and $(\hat{\mathbf{a}}_i, \hat{m}_i, \hat{r}_i, \hat{d}_i, \hat{\mathbf{n}}_i)$ be the corresponding renderings from the predicted scene. The overall training objective is

$$\mathcal{L} = \mathcal{L}_{\mathbf{a}} + \mathcal{L}_m + \mathcal{L}_r + \mathcal{L}_d + \mathcal{L}_{\hat{d}} + \mathcal{L}_{\mathbf{n}} + \mathcal{L}_{\hat{\mathbf{n}}} + \mathcal{L}_{\mathrm{rel}}. \tag{3}$$

Material supervision. We supervise intrinsic components using a combination of pixel-wise and perceptual losses. For each $X \in \{\mathbf{a}, m, r\}$, we define

$$\mathcal{L}_X = \sum_{i=1}^{N} \Big(\|\hat{X}_i - X_i\|_1 + \lambda_{\mathrm{lpips}}\, \mathrm{LPIPS}(\hat{X}_i, X_i) \Big). \tag{4}$$

Affine-invariant depth loss. As our training dataset provides depth maps, but doesn't contain camera parameters, we supervise rendered depth and predicted depth using an affine-invariant loss, inspired by MoGe (Wang et al., 2025b) for self-supervised depth estimation:

$$\mathcal{L}_d = \sum_{i=1}^{N} \frac{1}{|\Omega_i|} \sum_{p \in \Omega_i} \big(\delta_i(p) - \bar{\delta}_i\big)^2\,, \tag{5}$$

where $\delta_i(p) = \log \hat{d}_i(p) - \log d_i(p)$, $\bar{\delta}_i = \frac{1}{|\Omega_i|} \sum_{p \in \Omega_i} \delta_i(p)$, and Ω_i denotes the set of valid pixels.

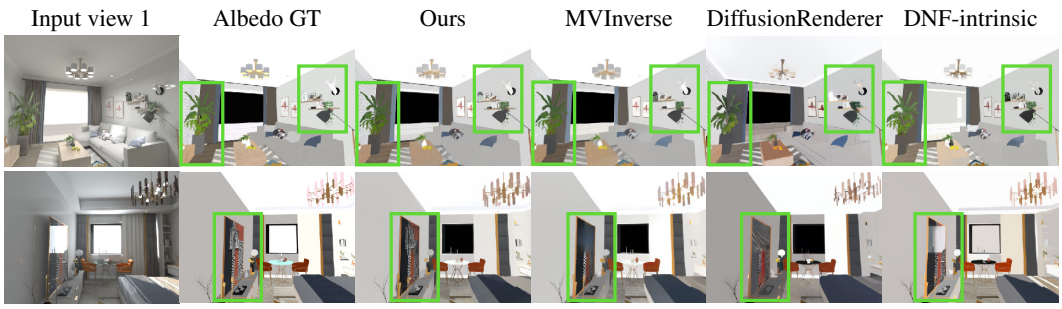

Figure 3: Qualitative reconstruction results on InteriorVerse from 2 input views. InvSplat jointly predicts geometry and intrinsic material attributes (albedo, metallic, roughness) in a single feed-forward pass, producing multi-view consistent decompositions.

Normal loss. Rendered normals and normals from the depth are supervised using a cosine similarity loss:

$$\mathcal{L}_n = \sum_{i=1}^{N} \left(1 - \langle \hat{\mathbf{n}}_i, \mathbf{n}_i \rangle\right). \tag{6}$$

Point-light relighting loss. Supervising materials and normals independently does not ensure that their combination produces correct shading under novel illumination. We therefore apply point-light relighting supervision to the Gaussian-rendered materials and normals using the shading model of Sec. 3.3. Unlike inference-time relighting, we use ground-truth depth to reconstruct surface positions $\mathbf{x}_p \leftarrow \mathbf{x}_p^{\mathrm{GT}}$, preventing errors in reconstructed geometry from affecting light directions and distances. We also omit cast shadows by setting $V_k(p) \leftarrow 1$ in Eq. (2), since ground-truth depth does not capture all occluders, while visibility from the predicted geometry changes abruptly under small changes of the Gaussians, which destabilizes training. This isolates the shading loss from errors in surface positions and visibility, while geometry is supervised separately. For each view, we sample K point lights near surface and shade each separately to provide supervision across lighting directions. Let $\hat{I}_k$ and I_k^{GT} denote the images shaded with Gaussian-rendered and ground-truth materials and normals, respectively. We minimize.

$$\mathcal{L}_{\mathrm{rel}} = \frac{1}{K} \sum_{k=1}^{K} \left\| \hat{I}_k - I_k^{\mathrm{GT}} \right\|_1, \tag{7}$$

where the image L_1 error is averaged over pixels and colour channels. Both images share the same surface positions, light, exposure, ambient-light formulation, and sRGB conversion. App. B illustrates this comparison alongside inference-time relighting.

4 Experiments

Implementation details. We train on curated image triplets from InteriorVerse (Zhu et al., 2022) at a resolution of 512×384. Two views serve as input, while ground-truth depth, normals, and material maps from all three views provide supervision. We initialize the geometry branch from ReSplat (Xu et al., 2025a) and the intrinsic branch and material and normal GS heads from MVInverse (Wu et al., 2025). Training requires approximately 12 hours for 20k steps with a batch size of 2 on a single H100 GPU.

Baselines. We compare against three 2D baselines that estimate material maps for the input views: DiffusionRenderer (Liang et al., 2025), DNF-Intrinsic (Zheng et al., 2025), and MVInverse (Wu et al., 2025).Since MVInverse was originally trained on datasets with different material definitions, we additionally fine-tune it on InteriorVerse to ensure consistent material supervision. For material estimation, previous 3D approaches rely on per-scene optimisation, which requires dense input views. We further evaluate three per-scene optimisation baselines in App. D.

Datasets. We use the standard training and test splits of InteriorVerse (Zhu et al., 2022) and evaluate cross-dataset generalisation on Structured3D (Zheng et al., 2020). Further dataset details are provided in the supplementary material. We additionally present qualitative results on procedurally generated

Table 1: Material reconstruction on the InteriorVerse test set with two input views per scene. * denotes fine-tuning on InteriorVerse; N/A indicates that model does not predict depth.

Method	Type	Flow Reprojection			Reconstruction							
		Albedo	Metallic	Roughness	Albedo			Metallic	Roughness	Normal	Relight-D_{GT}	Relight-D_{pred}
		RMSE↓	RMSE↓	RMSE↓	PSNR↑	SSIM↑	LPIPS↓	RMSE↓	RMSE↓	Cos. Sim.↑	PSNR↑	PSNR↑
DiffusionRenderer	2D	0.1129	0.1321	0.1132	17.32	0.800	0.253	0.2971	0.2825	0.9468	19.11	18.11
DNF-Intrinsic	2D	0.0889	0.1395	0.1235	18.64	0.850	0.211	0.1884	0.2124	0.9261	20.22	19.44
MVInverse	2D	0.0666	0.0652	0.0529	21.83	0.867	0.217	0.1039	0.1252	**0.9654**	21.91	N/A
MVInverse*	2D	0.0671	0.0628	0.0523	**22.92**	**0.886**	**0.182**	**0.0985**	**0.1221**	0.9630	21.44	N/A
Ours	**3D**	**0.0634**	**0.0533**	**0.0473**	22.23	0.875	0.211	0.0992	0.1268	0.9619	**22.36**	**21.63**

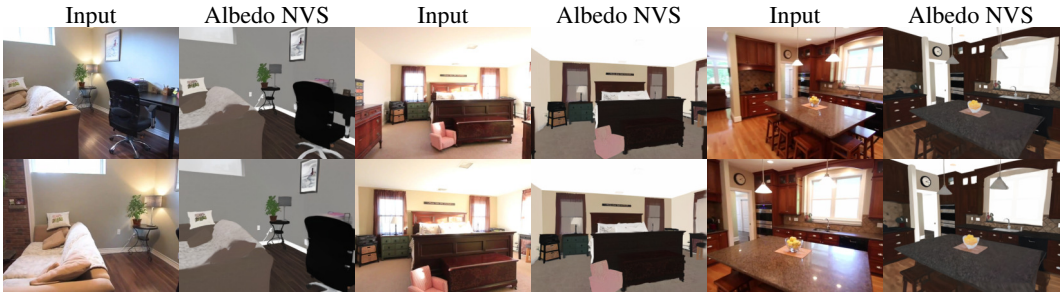

Figure 4: Generalization to real-world scenes from RealEstate10K. For each of the three scenes, we show the two input views and albedo renderings at two novel views produced by our model.

scenes from Infinigen (Raistrick et al., 2024) and real-world images from RealEstate10K (Zhou et al., 2018) and DL3DV (Ling et al., 2024).

Metrics. Following prior work (Wu et al., 2025; Zheng et al., 2025), we evaluate per-view material **reconstruction** on the input views using PSNR, SSIM, LPIPS, RMSE, and normal estimation using cosine similarity. Since albedo is ambiguous up to scale, we align each predicted albedo to the ground truth with a per-channel scale factor for all methods. We measure **cross-view consistency** by warping predictions between views using MegaFlow (Zhang et al., 2026) and computing reprojection RMSE for materials. **Relighting** quality is measured by PSNR between images shaded with predicted and ground-truth materials and normals under ten identical point lights per test view. **Relight-D_{GT}** uses ground-truth depth for both images, measuring the combined effect of material and normal errors on shading independently of depth reconstruction accuracy. **Relight-D_{pred}** instead uses each method's own predicted depth (Gaussian-rendered depth for ours and per-view depth maps for 2D methods), affinely aligned to ground truth per view, providing a relighting evaluation closer to inference-time conditions where ground truth depth is unavailable.

4.1 Evaluation

We first evaluate material and normals **reconstruction** on InteriorVerse. While image-space baselines predict pixel-aligned material maps, InvSplat reconstructs a shared 3D representation from which materials can be rendered into arbitrary views. As shown in Table 1, InvSplat achieves material and normal accuracy comparable to the strongest 2D baselines. Its main advantage is cross-view consistency: InvSplat obtains the lowest reprojection error for all three material channels, with the largest improvements in metallic and roughness. Qualitatively, InvSplat better separates specular reflections from albedo and preserves fine details (Fig. 3). App. C further illustrates how multi-view reconstruction helps resolve view-inconsistent appearance.

Beyond individual material metrics, we evaluate whether the predicted materials and normals produce consistent shading. InvSplat achieves the highest relighting PSNR with both ground-truth depth and predicted depth. The decrease with predicted depth reflects the additional effect of reconstruction errors on shading.

The two-view consistency evaluation can understate the benefits of a 3D representation, as rendering artefacts from sparse Gaussian reconstruction affect image-space metrics. These benefits become

Table 4: Ablation of the training objectives on InteriorVerse. Relight-D_{GT} reports PSNR between $\hat{I}_k$ and I_k^{GT}, both shaded using ground-truth depth. Relight-D_{pred} instead uses the predicted (rendered) depth for the prediction.

Variant	Albedo			Metallic	Roughness	Normal	Depth	Relight-D_{GT}	Relight-D_{pred}	Shadow
	PSNR↑	SSIM↑	LPIPS↓	RMSE↓	RMSE↓	Cos. Sim.↑	AbsRel↓	PSNR↑	PSNR↑	L1↓
Full model, K=5	**22.23**	**0.875**	0.211	**0.0992**	0.1268	**0.9619**	0.0461	**22.36**	**21.63**	0.0148
Full model, K=1	22.14	0.874	0.212	0.1001	0.1271	0.9614	0.0471	22.34	21.60	0.0144
w/o $\mathcal{L}_{\mathrm{rel}}$	22.00	0.868	0.211	0.1006	0.1267	0.9603	**0.0453**	21.71	21.12	**0.0092**
w/o $\mathcal{L}_{\mathrm{rel}}$, $\mathcal{L}_{\hat{d}}$ and $\mathcal{L}_{\hat{\mathbf{n}}}$	22.18	0.873	**0.203**	0.0993	**0.1254**	0.9609	0.0469	21.77	21.08	0.0180



Figure 5: **Qualitative relighting comparison on InteriorVerse.** To compare fairly against 2D baselines without geometry prediction, all methods are shaded on the same ground-truth depth under an identical novel point light; differences therefore arise from predicted materials and normals.

more apparent over longer sequences, where 2D feed-forward methods exhibit noticeable temporal flickering.

To assess **temporal consistency**, we evaluate 32-frame sequences from the first 100 scenes of RealEstate10K. Our model reconstructs each scene from the first and last frames and renders the intermediate views, while the 2D baselines process all 32 frames simultaneously. Despite using fewer input views, our method achieves the lowest temporal reprojection error across all material channels (Tab. 2), reducing albedo error by a factor of 1.3–3.9 and metallic and roughness errors by approximately 1.5–2. Further details are provided in Sec. G.

Table 2: Temporal consistency on first 100 32-frame RealEstate10K sequences: flow-based reprojection RMSE between consecutive frames. Ours uses only the first and last frame; 2D baselines see all 32. * denotes fine-tuned on InteriorVerse.

Method	#Views	Flow Reproj. RMSE ↓		
		Albedo	Metallic	Rough.
DiffusionRenderer	32	0.025	0.026	0.017
MVInverse	32	0.078	0.025	0.015
MVInverse*	32	0.027	0.029	0.020
Ours	2	**0.020**	**0.015**	**0.010**

4.2 Ablations

Dual-branch fusion. To evaluate our fusion design, we construct two baselines that directly combine MVInverse and ReSplat by assigning predicted per-view materials to the reconstructed Gaussians. The first freezes MVInverse and fine-tunes ReSplat, while the second jointly fine-tunes both networks. As shown in Tab. 3, our architecture outperforms both variants across all material properties, supporting our design of exchanging complementary geometry and intrinsic features rather than simply combining the two networks, and eliminates redundant backbones.

Table 3: Ablation study of 3D model design on InteriorVerse dataset using 2 input views per scene. * denotes fine-tuned network.

Method	Type	Albedo	Metallic	Roughness	Normal
		RMSE↓	RMSE↓	RMSE↓	Cos. Sim.↑
MVInverse + Resplat*	3D	0.1011	0.1041	0.1291	0.9582
MVInverse* + Resplat*	3D	0.0901	0.1011	0.1283	0.9607
Ours	3D	**0.0883**	**0.0993**	**0.1254**	**0.9609**

Training objective. We ablate the relighting loss and intermediate geometry supervision in Tab. 4. Removing $\mathcal{L}$rel reduces Relight-DGT from 22.36 to 21.71 dB and Relight-D_{pred} from 21.63 to 21.12 dB, demonstrating improved relighting quality even when using reconstructed depth. More lights provide better output quality. Fig. 5 illustrates the effect of relighting supervision. We also experimented with using predicted depth in the relighting loss, but found that lights positioned using ground-truth depth can fall inside the predicted geometry during early training and cause instabilities. We therefore use ground-truth depth for relighting supervision and supervise geometry separately. To assess the reliability of inference-time shadow tracing, we additionally measure visibility error by placing a point light at the camera center. Since every visible surface point should be directly illuminated from this position, the ground-truth visibility equals to one for every valid pixel; we report



Figure 6: Relighting and material editing on Infinigen (top) and RealEstate10K (bottom): (1) input, (2) reconstruction under the GT lamp for Infinigen and a manually selected novel light for RealEstate10K, (3) colored-light relighting, (4) increased roughness, and (5) increased metallic.

the mean absolute error of the ray-traced visibility. Although we supervise rendered depth, this alone is insufficient to prevent erroneous occlusions. Removing the intermediate depth and depth-derived normal losses, in the absence of relighting supervision, increases shadow error from 0.0092 to 0.0180, highlighting their importance for cast-shadow accuracy.

Normal prediction. We further study the design choice of explicitly predicting Gaussian normals. Among several alternatives, we find that using a dedicated normal prediction head yields the best performance. Additional details are provided in Supp. E.

4.3 Real-world generalisation and novel-view synthesis

We demonstrate scene reconstruction and novel-view synthesis on real-world scenes from RealEstate10K (Zhou et al., 2018) and DL3DV (Ling et al., 2024). As shown in Fig. 4, InvSplat reconstructs plausible, view-consistent albedo from two input views and renders it at novel viewpoints on unseen real-world scenes. Although trained with two input views, our model also accepts more views at inference, as illustrated by the four-view DL3DV examples in Fig. 15. However, inaccurate camera poses or depth estimates can introduce artifacts, particularly in challenging multi-view settings.

4.4 Applications

Relighting and material editing. We demonstrate relighting and material editing using the predicted material properties with a point-light renderer described in Sec. 3.3. We evaluate these applications on synthetic scenes generated with Infinigen (Raistrick et al., 2024) and real-world scenes from RealEstate10K. As shown in Fig. 6, our representation supports changes in scene materials and lighting, producing plausible appearance variations.

Potential for feed-forward reconstruction. Existing RGB-based reconstruction methods struggle to reproduce view-dependent effects, often baking specular highlights into appearance or geometry. By explicitly modeling scene materials, our representation can reproduce these effects under specified illumination while supporting both RGB reconstruction and relighting within a feed-forward framework. We illustrate this potential on synthetic scenes with known lighting Fig. 7 and real-world scenes Fig. 13.

5 Conclusion

We present InvSplat, a feed-forward model for scene reconstruction with material estimation from multi-view images. By estimating material parameters (albedo, roughness, metallic) together with the geometry of each Gaussian primitive, our method achieves high-quality, view-consistent materials and enables novel-view synthesis of material maps. Experimental results on synthetic and real-world data demonstrate strong performance and support applications such as relighting and material editing.

AI USE STATEMENT

We used generative AI tools to assist with manuscript drafting and language editing. The authors reviewed and revised all AI-assisted text and checked its consistency with the methods and experimental results. We take responsibility for the final content, including its accuracy, originality, and conclusions.

ETHICS STATEMENT

InvSplat supports scene reconstruction, material estimation, and relighting from images, with potential applications in graphics and content creation. Reconstructed scenes may retain private information present in the input images, and edited or relit outputs could be presented misleadingly as unaltered observations. Responsible use should respect image permissions and privacy and clearly identify manipulated content. Predicted geometry and material properties are estimates and should not be treated as verified physical measurements.

REPRODUCIBILITY STATEMENT

We describe the model architecture, Gaussian representation, rendering formulation, and training objectives in the method section. The experimental section and supplementary material document the datasets, preprocessing, training configuration, evaluation protocols, and baseline comparisons.

REFERENCES

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*, pp. 2447–2456, 2019.

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*, pp. 12684–12694, 2021.

David Charatan, Sizhe Lester Li, Andrea Tagliasacchi, and Vincent Sitzmann. pixelsplat: 3d gaussian splats from image pairs for scalable generalizable 3d reconstruction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 19457–19467, 2024.

Hongze Chen, Zehong Lin, and Jun Zhang. Gi-gs: Global illumination decomposition on gaussian splatting for inverse rendering. In *ICLR*, 2025.

Yuedong Chen, Haofei Xu, Chuanxia Zheng, Bohan Zhuang, Marc Pollefeys, Andreas Geiger, Tat-Jen Cham, and Jianfei Cai. Mvsplat: Efficient 3d gaussian splatting from sparse multi-view images. In *European conference on computer vision*, pp. 370–386. Springer, 2024.

Robert L. Cook and Kenneth E. Torrance. A reflectance model for computer graphics. *ACM Transactions on Graphics (TOG)*, 1:7 – 24, 1982. URL https://api.semanticscholar.org/CorpusID:278526816.

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*, pp. 73–89. Springer, 2024.

Kaiming He, Xiangyu Zhang, Shaoquing Ren, and Jian Sun. Deep residual learning for image recognition. 2016 ieee conf. In *Comput. Vis. Pattern Recognit*, pp. 770–778, 2015.

Lihan Jiang, Yucheng Mao, Linning Xu, Tao Lu, Kerui Ren, Yichen Jin, Xudong Xu, Mulin Yu, Jiangmiao Pang, Feng Zhao, Dahua Lin, and Bo Dai. Anysplat: Feed-forward 3d gaussian splatting from unconstrained views. *ACM Transactions on Graphics (TOG)*, 44:1 – 16, 2025a. URL https://api.semanticscholar.org/CorpusID:278997106.

Yu Jiang, Jiahao Xia, Jiongming Qin, Yusen Wang, Tuo Cao, and Chun-Xia Xiao. Humanmaterial: Human material estimation from a single image via progressive training. *ArXiv*, abs/2507.18385, 2025b. URL https://api.semanticscholar.org/CorpusID:280018546.

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

Peter Kocsis, Lukas Höllein, and Matthias Nießner. Intrinsic image fusion for multi-view 3d material reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026. URL https://peter-kocsis.github.io/IntrinsicImageFusion/.

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*, pp. 9039–9048, 2018.

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*, pp. 2475–2484, 2020.

Ruofan Liang, Zan Gojcic, Huan Ling, Jacob Munkberg, Jon Hasselgren, Zhi-Hao Lin, Jun Gao, Alexander Keller, Nandita Vijaykumar, Sanja Fidler, and Zian Wang. Diffusionrenderer: Neural inverse and forward rendering with video diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2025. URL https://research.nvidia.com/labs/toronto-ai/DiffusionRenderer/.

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*, pp. 21644–21653, 2024.

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 *CVPR*, 2025.

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*, pp. 22160–22169, 2024.

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. doi: 10.48550/arXiv.2510.10726. URL https://arxiv.org/abs/2510.10726.

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.

Steven G Parker, James Bigler, Andreas Dietrich, Heiko Friedrich, Jared Hoberock, David Luebke, David McAllister, Morgan McGuire, Keith Morley, Austin Robison, et al. Optix: a general purpose ray tracing engine. *Acm transactions on graphics (tog)*, 29(4):1–13, 2010.

Siddhant Prakash, Gilles Rainer, Adrien Bousseau, and George Drettakis. Deep scene-scale material estimation from multi-view indoor captures. *ArXiv*, abs/2211.08047, 2022. URL https://api.semanticscholar.org/CorpusID:252822753.

Alexander Raistrick, Lingjie Mei, Karhan Kayan, David Yan, Yiming Zuo, Beining Han, Hongyu Wen, Meenal Parakh, Stamatis Alexandropoulos, Lahav Lipson, Zeyu Ma, and Jia Deng. Infinigen indoors: Photorealistic indoor scenes using procedural generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21783–21794, June 2024.

René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2020.

Katharina Schmid, Nicolas von Lützow, Jozef Hladký, Angela Dai, and Matthias Nießner. Genrecon: Bridging generative priors for multi-view 3d scene reconstruction. *ArXiv*, abs/2605.23888, 2026. URL https://api.semanticscholar.org/CorpusID:288649325.

Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4104–4113, 2016.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025a.

Ruicheng Wang, Sicheng Xu, Cassie Dai, Jianfeng Xiang, Yu Deng, Xin Tong, and Jiaolong Yang. Moge: Unlocking accurate monocular geometry estimation for open-domain images with optimal training supervision. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 5261–5271, 2025b.

Xiangzuo Wu, Chengwei Ren, Jun Zhou, Xiu Li, and Yuan Liu. Mvinverse: Feed-forward multi-view inverse rendering in seconds. *arXiv preprint arXiv:2512.21003*, 2025. URL https://maddog241.github.io/mvinverse-page/.

Haofei Xu, Daniel Barath, Andreas Geiger, and Marc Pollefeys. Resplat: Learning recurrent gaussian splats. *arXiv preprint arXiv:2510.08575*, 2025a.

Haofei Xu, Songyou Peng, Fangjinhua Wang, Hermann Blum, Daniel Barath, Andreas Geiger, and Marc Pollefeys. Depthsplat: Connecting gaussian splatting and depth. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 16453–16463, 2025b.

Dingxi Zhang, Fangjinhua Wang, Marc Pollefeys, and Haofei Xu. Megaflow: Zero-shot large displacement optical flow. In *European Conference on Computer Vision (ECCV)*, 2026.

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.

Hengshuang Zhao, Li Jiang, Jiaya Jia, Philip HS Torr, and Vladlen Koltun. Point transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 16259–16268, 2021.

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*, pp. 519–535. Springer, 2020.

Rongjia Zheng, Qing Zhang, Chengjiang Long, and Wei-Shi Zheng. Dnf-intrinsic: Deterministic noise-free diffusion for indoor inverse rendering. *arXiv preprint arXiv:2507.03924*, 2025. doi: 10.48550/arXiv.2507.03924. URL https://arxiv.org/abs/2507.03924v2. Accepted to ICCV 2025.

Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. *arXiv preprint arXiv:1805.09817*, 2018.

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*, pp. 1–8, 2022.

CONTENTS

A View-dependent effects modelling 13
B Relighting during training and inference 14
C Resolving view-inconsistent appearance 14
D Comparison with optimization-based inverse rendering 15
E Normal ablation. 16
F Structured3D evaluation 16
G Temporal consistency on video sequences 16
H Implementation Details 18
I Limitations 18
J Real-world generalizability 19

A VIEW-DEPENDENT EFFECTS MODELLING OF FEED-FORWARD SCENE RECONSTRUCTION METHODS

When ground-truth lighting is available (as in our Infinigen scenes), this allows the renderer to reproduce view-dependent effects by construction, leading to more faithful relit RGB renderings and improved consistency across novel viewpoints. We provide an example on synthetic scene generated with Infinigen in Figure 7. In this example, we knew the light configuration and used it to render the reconstructed Gaussians, which allows us to faithfully reproduce view-dependent effects such as specular highlights. In contrast, AnySplat and WorldMirror, which predict RGB appearance directly, struggle to capture these effects and tend to bake them into color. ReSplat has similar effect, although it it less visible than on other methods. Additionally, we show a similar effect on real-world scenes in Figure 13, where we don't have access to ground-truth lighting. In this case, we used a point light source and placed it in the middle of the scene. Although our model is not perfectly reproduce the original RGB image, we can see realistic highlight movements, while other methods struggle to capture them. This shows that our model can learn to separate view-dependent effects from albedo, producing more realistic relighting results.

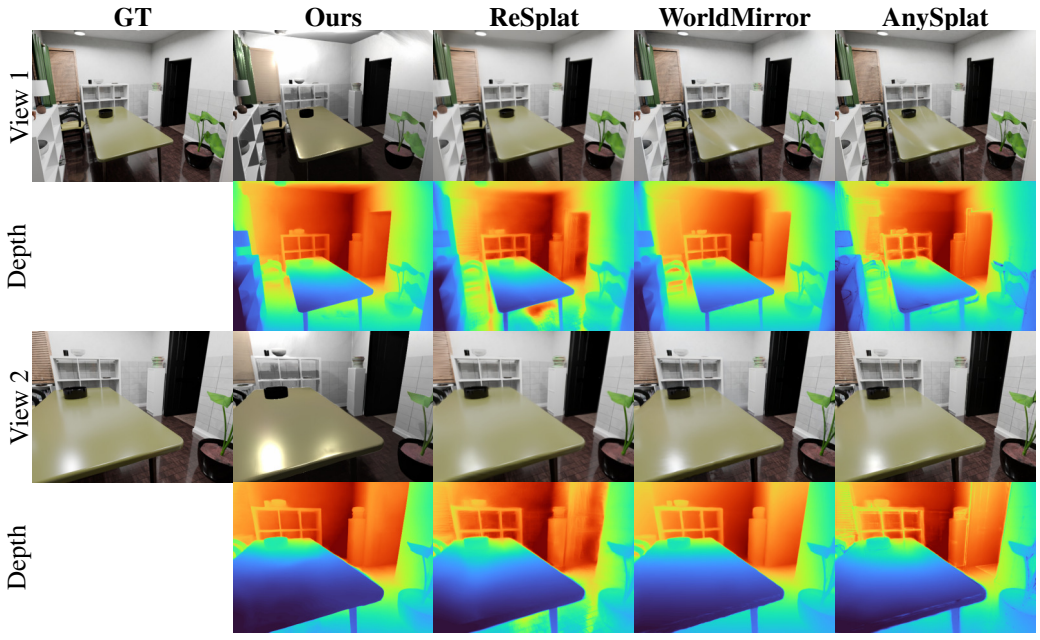

Figure 7: RGB reconstruction comparison on a synthetic Infinigen scene. Our method renders RGB by composing predicted material factors with the ground-truth Infinigen light configuration. For classical feed-forward RGB reconstruction methods, without explicitly modeling material, the view-dependent effect is limited.

B RELIGHTING DURING TRAINING AND INFERENCE

Fig. 8 illustrates the relationship between the relighting loss, the evaluation metric, and the inference renderer described in Sec. 3.3. The reference I_k^{GT} and prediction $\hat{I}_k$ are both shaded using ground-truth depth without cast shadows; they differ only in their materials and normals. This is the image pair compared by the training loss using an L_1 error and by **Relight-D_{GT}** using PSNR. Here, D_{GT} indicates that ground-truth depth is shared by the reference and prediction, allowing the metric to assess material and normal quality independently of depth reconstruction errors.

The inference render I_k instead uses the reconstructed depth and ray-traced light visibility. All three images share the same point-light position, intensity, and exposure fitted to the reference. While $\hat{I}_k$ closely reproduces the reference shading, I_k additionally shows cast shadows from the chandelier, floor lamp, and furniture. The figure thus connects the controlled comparison used for supervision and evaluation to relighting with the reconstructed 3D scene.



Figure 8: **Relighting comparison.** Left to right: two input views, reference I_k^{GT}, prediction $\hat{I}_k$, and inference render I_k with cast shadows. Relight-D_{GT} measures PSNR between I_k^{GT} and $\hat{I}_k$. The white sphere marks the point light.

C RESOLVING VIEW-INCONSISTENT APPEARANCE

Our model predicts pixel-aligned Gaussians for each input view and uses the predicted depth to unproject them into a shared world coordinate frame. The scene reconstruction is the union of these per-view Gaussian sets. To examine their individual contributions, we render albedo from the combined set and from each subset separately, keeping the camera and all predicted Gaussian parameters fixed (Fig. 9). Each subset is taken from the same joint prediction, rather than reconstructed from a single image.



Figure 9: Per-view contributions to the reconstructed albedo. Left to right: the two input RGB images, followed by albedo rendered using all Gaussians, only those associated with view 1, and only those associated with view 2. All three albedo panels use the camera of input view 1 and a black background.

The shared 3D representation gives the model flexibility to resolve inconsistent observations caused by specular highlights and reflections: it can suppress contributions from affected views through the predicted Gaussian parameters and recover surface texture from other views. For example, the window reflection obscures the painting in view 1, while the combined reconstruction recovers the canvas colours visible in view 2. This illustrates how the model can reconcile conflicting appearance cues within a single albedo reconstruction.

Table 5: Comparison with per-scene optimization-based inverse rendering methods (GS-IR, GI-GS, Intrinsic Image Fusion (IIF)) on 16 InteriorVerse test scenes (3–46 frames per scene). Optimization-based baselines are fitted on *all* frames of each scene, whereas our method uses only 2 input views; all methods are evaluated on the same 2 frames. Reconstruction time is measured per scene.

Method	**Type**	**Albedo**			**Metallic**	**Roughness**	**Time↓**
		PSNR↑	**SSIM↑**	**LPIPS↓**	**RMSE↓**	**RMSE↓**	
GI-GS (Chen et al., 2025)	Optim.	10.75	0.537	0.555	0.342	0.632	2 h
GS-IR (Liang et al., 2024)	Optim.	12.35	0.659	0.440	0.319	0.609	1 h
IIF (Kocsis et al., 2026)	Optim.	12.62	0.629	0.448	0.222	0.408	0.5 h
Ours	Feed-forward	**23.22**	**0.887**	**0.173**	**0.077**	**0.105**	**1.5 s**

D Comparison with optimization-based inverse rendering

Our main comparison focuses on feed-forward methods, reflecting our goal of efficient scene-level inverse rendering from sparse views. Here, we additionally evaluate three per-scene optimisation methods: GS-IR (Liang et al., 2024) and GI-GS (Chen et al., 2025), which optimise 3D Gaussians using physically based shading and an environment map, and Intrinsic Image Fusion (IIF) (Kocsis et al., 2026), which fuses per-view intrinsic priors into a consistent material representation and refines it through inverse path tracing.

Protocol. Since optimisation-based methods typically require dense observations, we select 10 InteriorVerse test scenes with successfully estimated COLMAP cameras and relatively many frames (22–46 per scene), together with 6 randomly selected sparse-view scenes (3–15 frames). Each baseline is fitted to all available frames, including the evaluation views, whereas our method uses only the two input views specified by the standard test protocol. We evaluate all methods on these same two views using the main-paper metrics, including per-channel scale alignment for albedo. We use the official baseline implementations with default hyperparameters and render their optimised albedo, metallic, and roughness maps into the evaluation views. Reconstruction times are measured per scene on a single A100 GPU.

Results. As shown in Table 5, our method achieves substantially higher material accuracy despite using fewer input views. Albedo PSNR reaches 23.22 dB, compared with 10.75 dB for GI-GS, 12.35 dB for GS-IR, and 12.62 dB for IIF. Our metallic and roughness RMSE are 0.077 and 0.105, respectively, compared with 0.22–0.34 and 0.41–0.63 for the optimisation baselines. Reconstruction takes 1.5 s per scene, versus 0.5–2 h, a speedup of more than three orders of magnitude.

GS-IR and GI-GS primarily optimise RGB reconstruction without explicit material supervision. Their global environment lighting model may also inadequately capture local emitters and inter-reflections in indoor scenes, contributing to shadows and highlights being absorbed into albedo and ambiguity in metallic and roughness estimates. IIF incorporates 2D material priors and achieves higher albedo PSNR than these methods, but remains below our directly supervised model. These results indicate that additional views alone do not ensure accurate material decomposition and highlight the importance of suitable priors and training objectives.

Complementarity with optimisation. The explicit geometry and per-Gaussian materials predicted by InvSplat could also initialise subsequent inverse rendering. For example, lighting could be estimated from the reconstructed scene rather than jointly optimising geometry, materials, and illumination from scratch. We leave this extension to future work.

E NORMAL ABLATION.

We also ablate the normal prediction branch and test two other variants: one in which we directly compute normals from depth using finite differences, and another in which we predict normals from the Gaussian head together with scales and opacities and then render them. In all experiments, we used a normal consistency loss with the ground truth. In both variants, the model struggles to separate geometry from appearance and leaks texture details into the normals, which harms relighting quality. Results are shown in Figure 10. We additionally tried adding a normal smoothness loss, but this did not help much. This shows that pre-training also helps when learning Gaussian normals jointly with material factors.

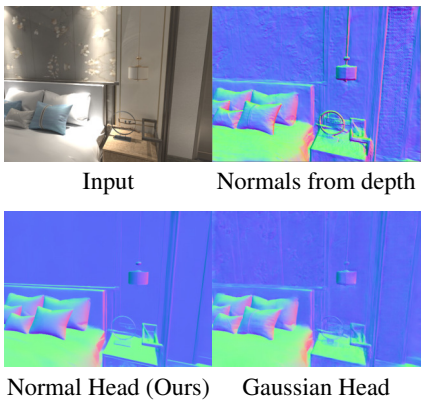

Figure 10: Alternative derivation for gaussian normals. First row: input image, normals derived from depth. Second row: rendered normals, left is separate head for prediction, right is prediction in gaussian head.

F STRUCTURED3D EVALUATION

For **cross-view consistency** evaluation, we use the Structured3D dataset, which provides ground-truth albedo maps, depth maps, and camera parameters for each scene. Since the dataset does not contain smooth camera trajectories, we apply a depth-reprojection-based view selection strategy with an overlap threshold of 0.5, and evaluate on the first 251 scenes. Quantitative comparisons are reported in Table 6.

Table 6: Multi-view consistency and albedo reconstruction on Structured3D using 2 input views per scene. * denotes fine-tuned model (on InteriorVerse). Our method achieves better multi-view consistency, especially for metallic and roughness, with the same reconstruction quality.

Method	**Type**	**Reprojection RMSE ↓**			**Albedo Reconstruction**			
		Albedo	**Metallic**	**Roughness**	**PSNR↑**	**SSIM↑**	**LPIPS↓**	**RMSE↓**
DiffusionRenderer	2D	0.100	0.122	0.108	15.75	0.714	0.310	0.174
DNF-Intrinsic	2D	0.122	0.183	0.147	14.37	0.703	0.303	0.209
MVInverse	2D	0.044	0.056	0.038	19.83	0.771	0.268	0.108
MVInverse*	2D	**0.037**	0.051	0.034	**20.48**	**0.798**	**0.247**	**0.101**
Ours	3D	0.044	**0.048**	**0.030**	19.57	0.779	0.281	0.112

Additionally, Figure 11 visualizes the prediction from view 1 warped into view 0, together with the corresponding error maps. Existing 2D-based methods exhibit noticeable inconsistencies across views, particularly around reflective surfaces and bright highlights. In contrast, our method produces more consistent predictions due to its unified 3D Gaussian representation. The improvement is especially pronounced for metallic and roughness, which are more challenging to infer from RGB images and benefit from stronger multi-view aggregation.

We further evaluate albedo reconstruction quality, demonstrating that our model does not simply overfit to the InteriorVerse dataset, but instead learns a more generalizable material representation.

G TEMPORAL CONSISTENCY ON VIDEO SEQUENCES

The cross-view consistency evaluation on Interiorverse and Structured3D is limited to two input views per scene, since the dataset does not provide smooth camera trajectories. In this regime, rendering artifacts from sparse Gaussian reconstruction (e.g., at depth discontinuities or in regions observed by only one view) are counted as inconsistencies when the metric is computed in image space, which partially masks the advantage of an explicit 3D representation. To evaluate consistency in the practical setting of long image sequences, where 2D feed-forward methods exhibit noticeable

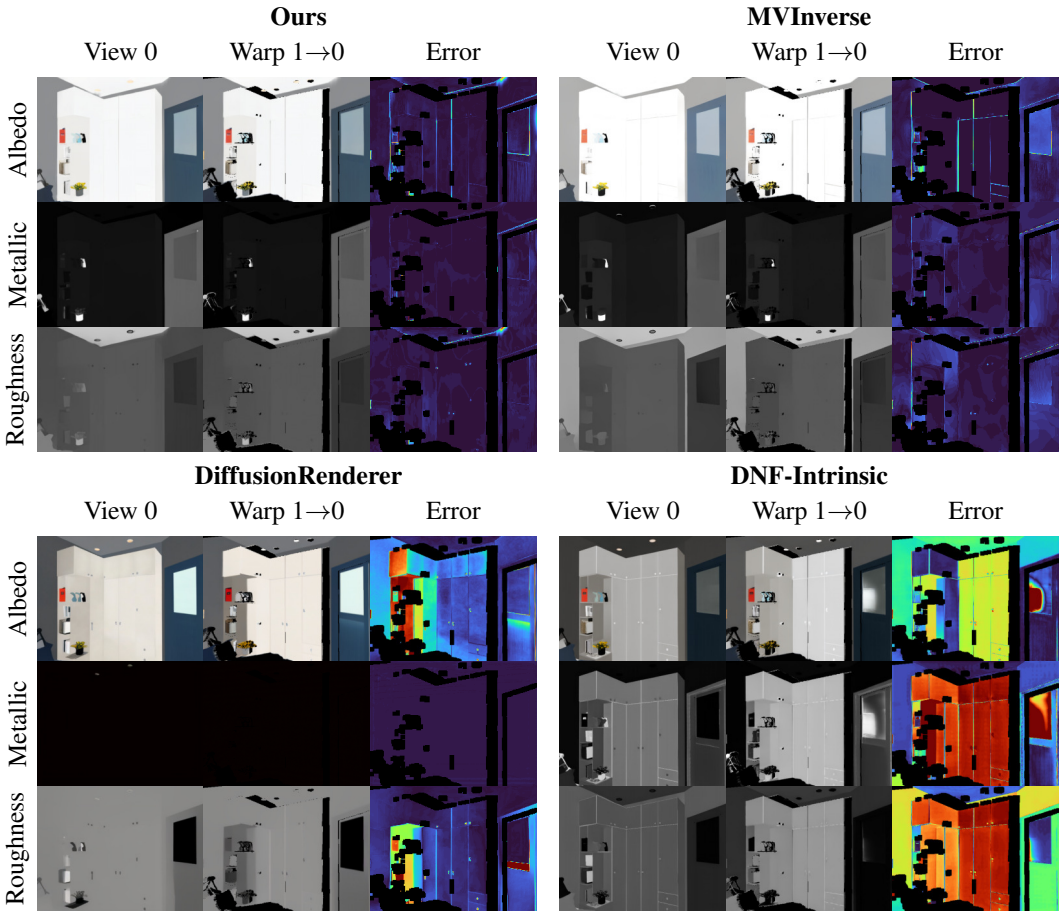

Figure 11: Multi-view material consistency on a scene from Structured3D. For each method, the figure shows prediction at view 0, the prediction from view 1 warped into view 0 using ground-truth depth and pose, and the per-pixel error between them. See Supp. Figure 18 for additional examples.

temporal flickering, we additionally evaluate on real-world videos from RealEstate10K (Zhou et al., 2018).

Protocol. We use the first 100 scenes of RealEstate10K test split and extract a 32-frame sequence from each. Our model reconstructs the scene from only the first and last frame (2 input views) and renders the material maps for all 32 frames, whereas the 2D baselines (DiffusionRenderer, MVInverse, and MVInverse fine-tuned on InteriorVerse) process all 32 frames simultaneously as input. Since RealEstate10K provides neither camera poses nor ground-truth geometry, we cannot use the depth-based reprojection metric of the main paper. Instead, we estimate dense frame-to-frame optical flow with MegaFlow (Zhang et al., 2026), warp the material prediction of frame t into frame $t{+}1$, and compute the RMSE against the prediction of frame $t{+}1$ over all valid correspondences. The error is averaged over all consecutive frame pairs and all scenes. Note that the flow is estimated on RGB frames and is therefore the same for all methods, so the metric only measures the consistency of the predicted materials.

Results. Table 2 reports the results. Our method achieves the lowest temporal error on all three material channels (albedo 0.020, metallic 0.015, roughness 0.010), compared to 0.025/0.026/0.017 for DiffusionRenderer, 0.078/0.025/0.015 for MVInverse, and 0.027/0.029/0.020 for the fine-tuned MVInverse. The gap is considerably larger than in the two-view Structured3D setting: while the 2D methods must re-estimate materials for every frame and thus produce flicker, in particular on reflective surfaces and around light sources, our method predicts a single 3D scene once and renders it consistently along the trajectory. Notably, this is achieved while observing only 2 of the 32 frames,

so our method also requires substantially fewer input views. Fine-tuning MVInverse on InteriorVerse improves its albedo consistency but slightly degrades metallic and roughness consistency, which indicates that image-space fine-tuning alone does not resolve the multi-view inconsistency of 2D methods.

H Implementation Details

Dataset. InteriorVerse provides ground-truth material maps (albedo/metallic/roughness) and geometry maps (depth/normal). Since it does not provide camera poses and views are sampled randomly within scenes (rather than along a smooth video trajectory), we first estimated camera poses using COLMAP (Schonberger & Frahm, 2016), and then selected views with large overlap (>0.4) via depth reprojection. As the dataset is very sparse (between 2 and 80 views per scene) and indoor scenes also contain large textureless regions, we found COLMAP estimation can be noisy and incomplete. To mitigate this, we ran ReSplat to select triplets with good RGB reconstruction quality (PSNR > 23) for training. For testing, we used one pair per scene with the largest overlap, without any ReSplat filtering, to avoid introducing selection bias when comparing to 2D methods. We followed the InteriorVerse train/test split. After all filtering, we obtained 13k triplets for training and 137 pairs for testing.

Following several of our baselines, as InteriorVerse images contain noise from dataset rendering, we run the OptiX (Parker et al., 2010) denoiser on the input RGB images and use tone-mapping. We found that each baseline performs better with either the original input or the denoised input, and we report the best results for each method.

Training details. We train with AdamW (weight decay 0.01) and gradient clipping at 0.5, using a linear warmup over 1000 steps followed by cosine annealing. The ResNet of the geometry branch is frozen. The DINOv2 encoder is fine-tuned with a learning rate of 10^{-6}, and all other components, including the multi-view intrinsic translator, with 10^{-5}. The loss weights are 1.0 for $\mathcal{L}_{\mathbf{a}}$, 0.5 for $\mathcal{L}_m$ and $\mathcal{L}_r$, $\lambda_{\text{lpips}} = 1.0$, 1.0 for $\mathcal{L}_d$, 0.1 for $\mathcal{L}_{\hat{d}}$, 1.0 for $\mathcal{L}_{\mathbf{n}}$ and $\mathcal{L}_{\hat{\mathbf{n}}}$, and 75 for $\mathcal{L}_{\text{rel}}$. Since metallic and roughness maps have a single channel, we replicate them to three channels to compute LPIPS. For the fine-tuned MVInverse baseline, we fine-tune the DINOv2 encoder, the translator, and the material and normal heads with a learning rate of 10^{-5} and 500 warmup steps.

Relighting loss. The BRDF f_{CT} is the sum of a Lambertian diffuse lobe $(1-m)\,\mathbf{a}/\pi$ and a microfacet specular lobe with the GGX normal distribution, the Smith–Schlick geometry term, and Schlick's Fresnel approximation with base reflectance $F_0 = 0.04\,(1-m) + m\,\mathbf{a}$. Roughness is clamped to $r \geq 0.05$ for numerical stability. For each input view, we render $K = 5$ image pairs at half resolution, each lit by a single white point light with intensity $\Phi = 10$. Each light lies on the camera ray of a pixel drawn at random from the central half of the image, at 0.8 of the ground-truth depth of that pixel, and is resampled at every iteration. Each image is formed as $\Gamma(S) = \text{sRGB}(e_i\,S + \epsilon\,\mathbf{a})$. The per-view exposure e_i maps the median luminance of the lit pixels of the ground-truth shading to 0.15. It is computed from the ground-truth image only, shared by both images, and receives no gradient. The exposure and the sRGB transfer function keep the error from being dominated by the few bright pixels close to the light. The ambient term, with $\epsilon = 0.02$, provides an albedo signal in pixels that the light does not reach, which would otherwise be black in both images, and stands in for the indirect illumination that a single point light does not produce.

The **Relight-D_{GT}** PSNR in Tab. 1 compares $\hat{I}_k$ with I_k^{GT} using the same shadow-free shading model as the relighting loss. Each test view is relit once with the predicted albedo, metallic, roughness, and normal maps of a method and once with the ground-truth maps, both on the ground-truth depth, under ten point lights placed inside the scene: each light lies on the camera ray of a pixel drawn at random from the central half of the image, at 0.8 of the ground-truth depth of that pixel. The lights are drawn once with a fixed seed from the ground-truth depth only and are therefore identical for all methods; we report the PSNR between the two relit images, averaged over the ten lights. As for the other albedo metrics, the predicted albedo is aligned to the ground truth with one scale per scene.

I Limitations



Figure 13: RGB reconstruction comparison on DL3DV across four views (rows) and methods (columns). Although we don't estimate scene lighting and can't produce exactly the same RGB reconstruction, our method can disentangle light from appearance and our highlights move with view change, while RGB-based methods struggle with this.

Our model inherits the limitations of 2 domains. From the feed-forward scene reconstruction side, if poses are estimated incorrectly, our reconstruction will produce corrupted results, as illustrated in Figure 12. Also, performance on more views and different resolutions is limited. Additionally, as our feed-forward model is non-generative, it lacks the ability to reconstruct unseen regions of a scene with detailed geometry and texture.



View 1 View 2

Figure 12: Failure case example.

Our feed-forward inverse rendering model significantly accelerates 3D content creation, democratizing asset generation for AR/VR while reducing the heavy energy consumption associated with traditional per-scene optimization. However, this increased accessibility also lowers the barrier for malicious applications, such as generating deceptive 3D media or unauthorized replication of intellectual property. Additionally, our model may inherit biases from its training data. Mitigating these risks requires ongoing community efforts in 3D provenance and diverse dataset curation.

J Real-world generalizability

We provide results of inverse rendering of our model on DL3DV dataset with 2 input views in Fig. 14 and with 4 input views in Fig. 15.

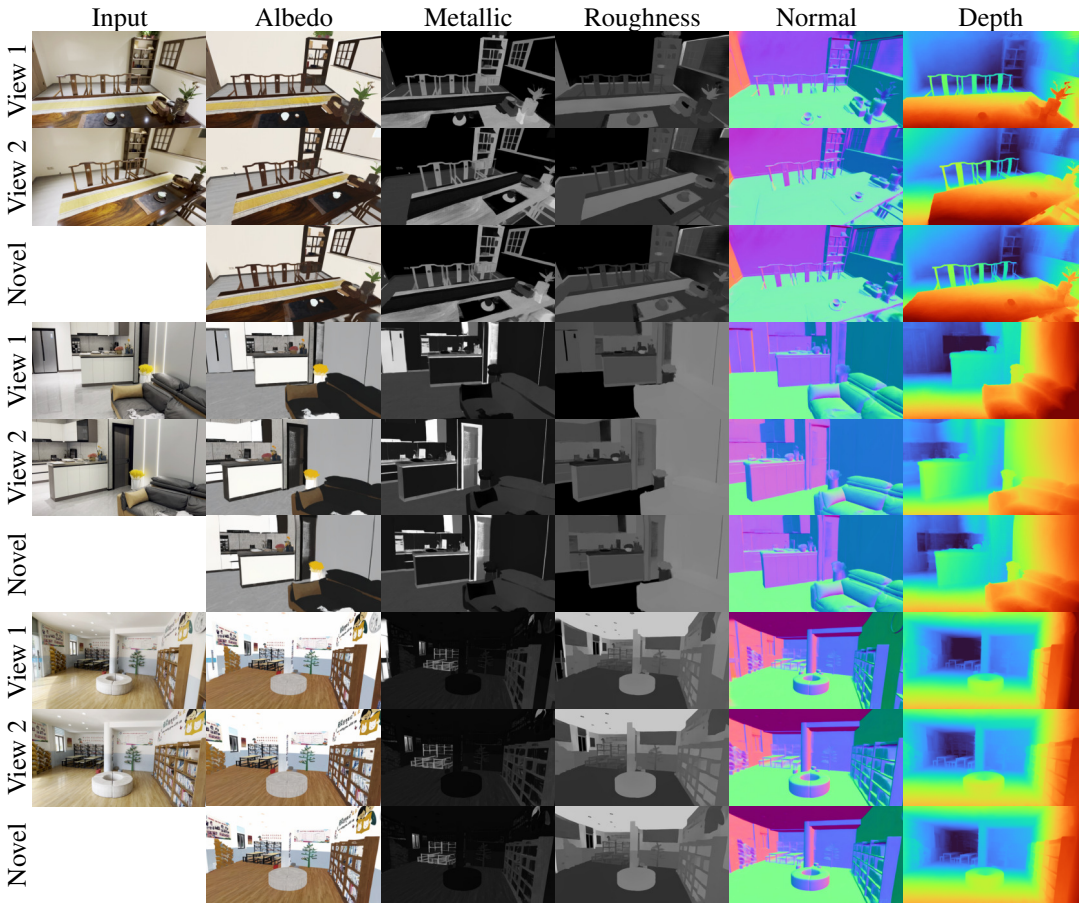

Figure 14: Qualitative results on three DL3DV scenes with 2 input views each. For every scene the first two rows show the predicted intrinsics (albedo, metallic, roughness, normal, depth) at the two input views. The third row shows novel-view synthesis results at an interpolated viewpoint between them.

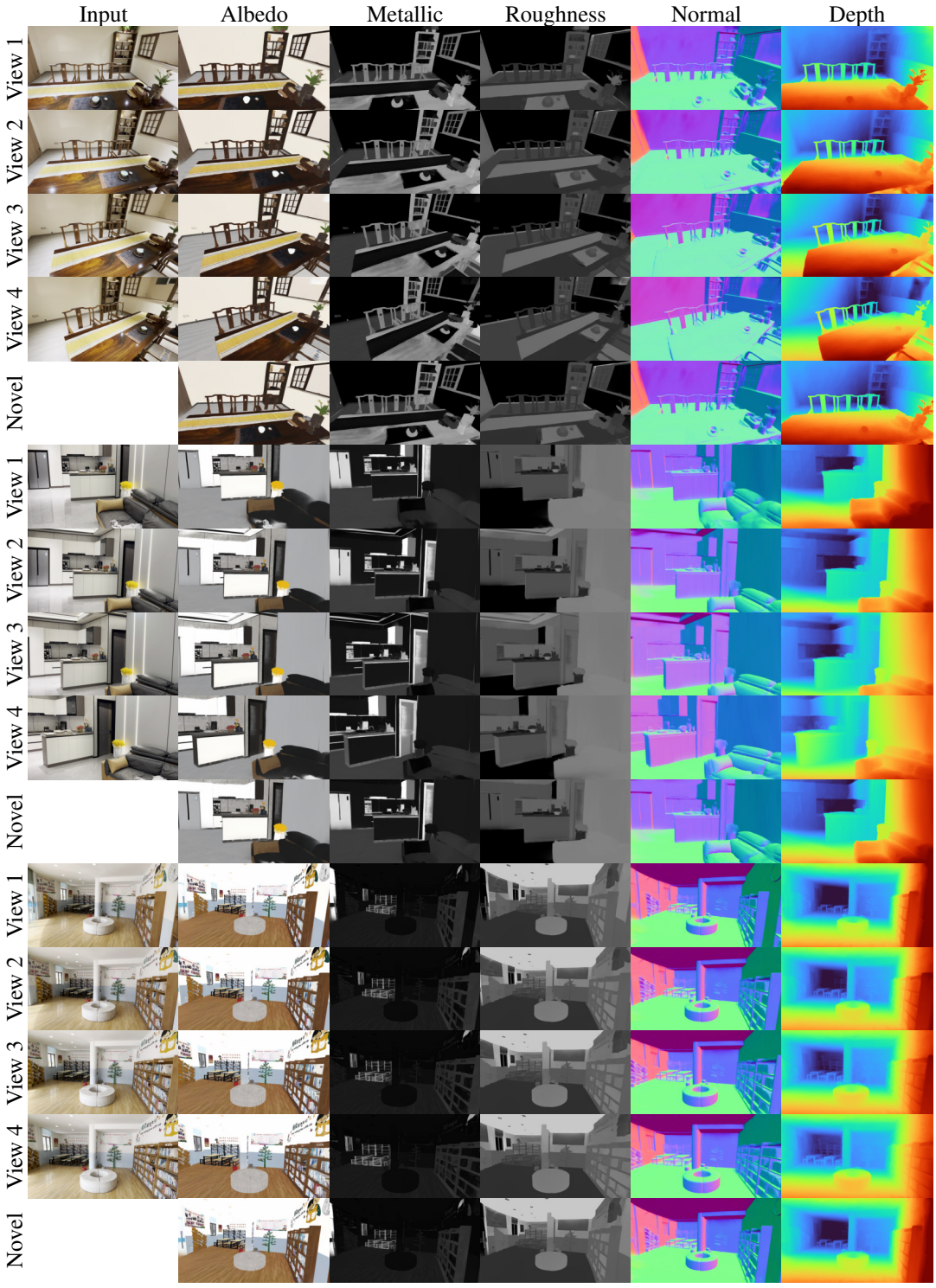

Figure 15: Qualitative results on three DL3DV scenes with 4 input views each. For every scene the first four rows show the predicted intrinsics (albedo, metallic, roughness, normal, depth) at the four input views. The fifth row shows novel-view synthesis results at an interpolated viewpoint between the first two input views.

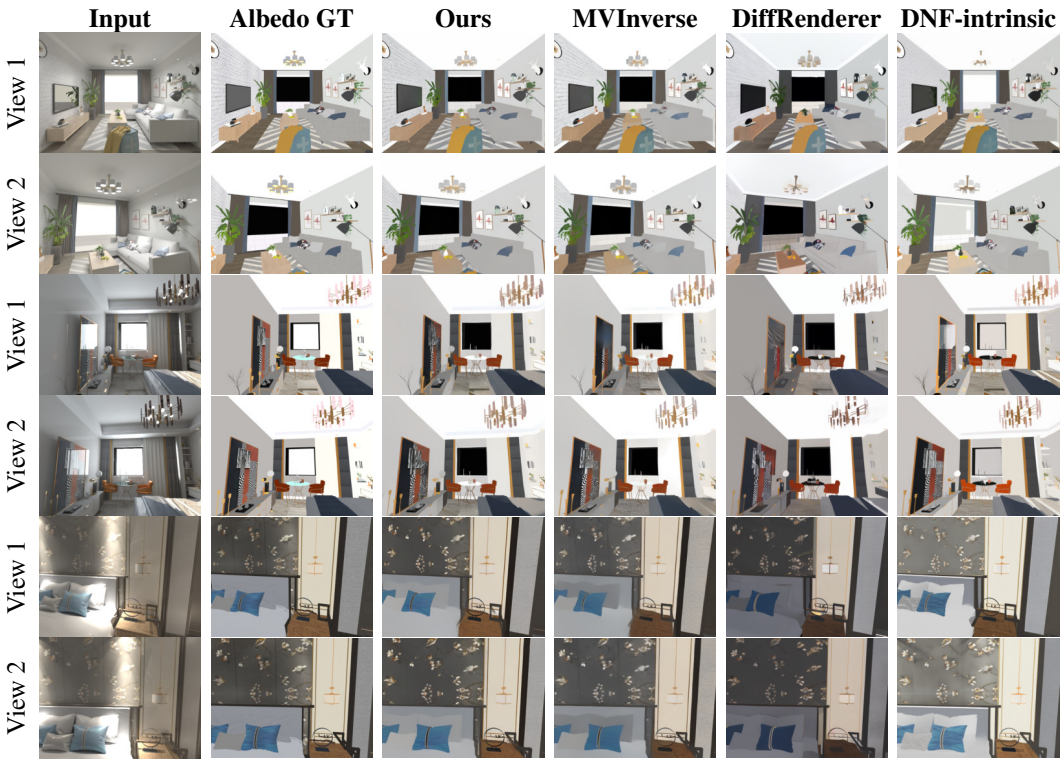

Figure 16: Qualitative albedo comparison on InteriorVerse. Each row shows one view of a scene. Columns: input RGB, albedo ground truth, and predicted albedo from different methods.

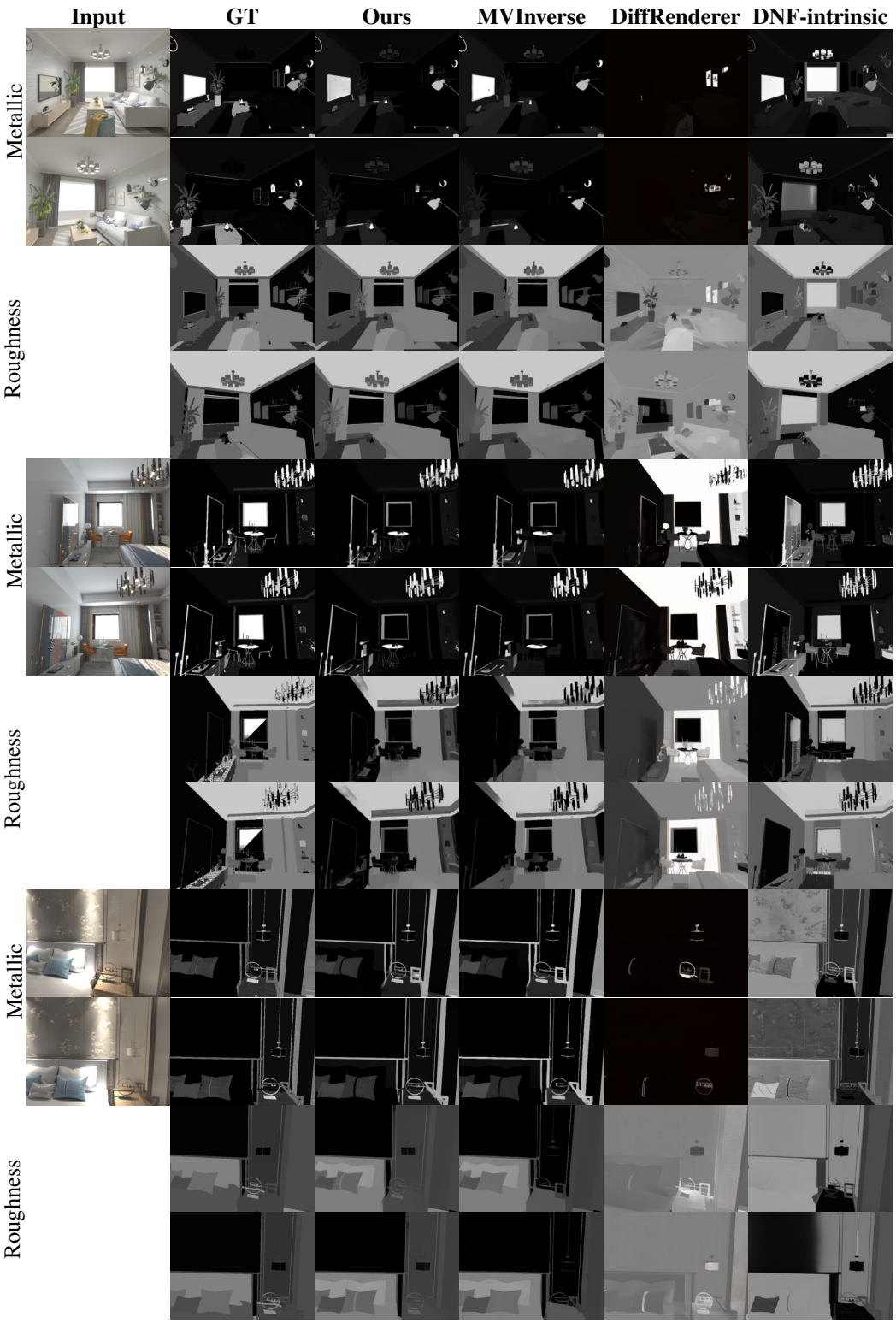

Figure 17: Qualitative metallic and roughness comparison on InteriorVerse for the same 3 scenes as Figure 16. For each scene we show 2 metallic rows followed by 2 roughness rows (one row per view). Columns: input RGB, ground truth, and predictions from each method.

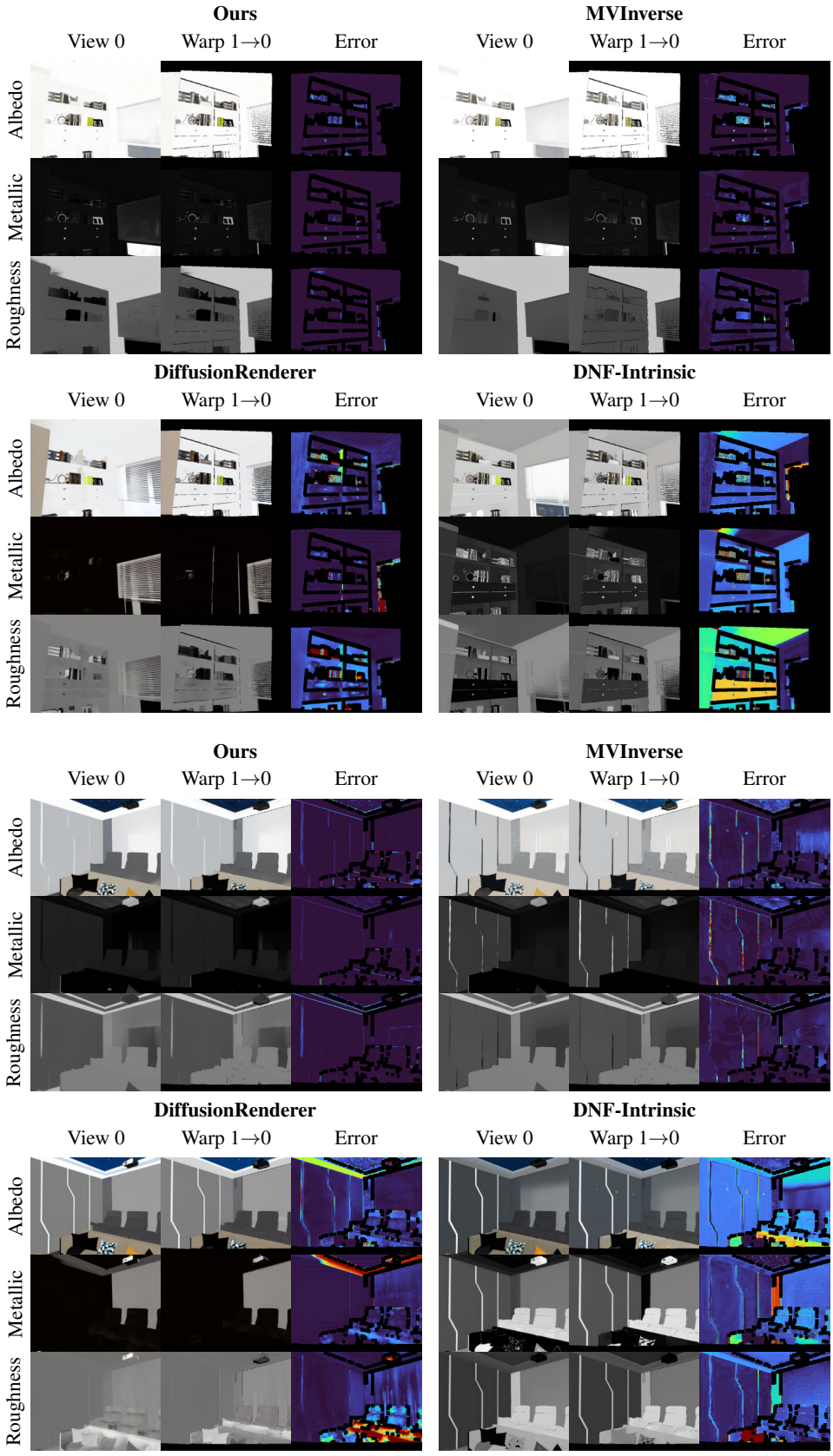

Figure 18: Multi-view consistency on additional examples from Structured3D for albedo, metallic and roughness. For each method, the figure shows the prediction at view 0, the prediction from view 1 warped into view 0 using ground-truth depth and pose, and the per-pixel error between them.