# Dynamic Inverse Rendering for Enhanced Material-Lighting Decomposition

Raza Yunus[1,3], Benjamin Ummenhofer[2], Jan Eric Lenssen[3], and Eddy Ilg[1,†]

[1] University of Technology Nuremberg
[2] Intel
[3] Max Planck Institute for Informatics, Saarland Informatics Campus
Project Page: razayunus.github.io/DIR. [†] Now at Google.

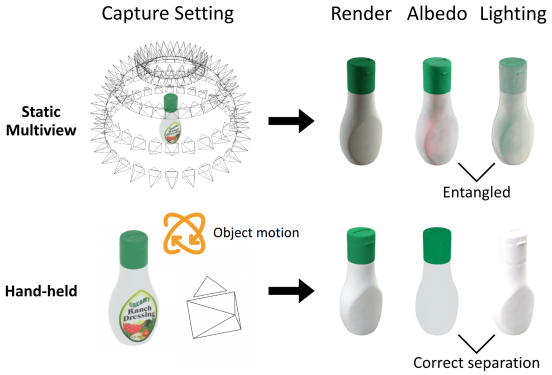


**Fig. 1:** In contrast to the static multiview setting, where albedo and the corresponding lighting often become entangled, we use the hand-held object capture setting in which diverse surface-light interactions impose strong constraints on the optimization and enable proper material-lighting disentanglement.

**Abstract.** Decomposing outgoing surface radiance into material and illumination during inverse rendering is essential for applications such as relighting and augmented reality, yet it is severely ill-posed since multiple combinations can result in the same observed colour. Capturing an object under multiple lighting conditions usually helps resolve this ambiguity as it constrains the optimization towards correct solutions. In this work, we explore the potential of reconstructing rigidly moving objects—which provides observations of diverse light-surface interactions—to resolve the material-lighting ambiguity in inverse rendering. For this purpose, we introduce a relightable approach that marries object tracking and reconstruction with inverse rendering for general rigidly moving objects. Our experimental analysis on synthetic data demonstrates that motion can be an advantage for disentangling material and lighting: the reconstructed material is significantly more accurate when the object is observed under rigid motion than when it is static. Moreover, results on RGB videos of real hand-held objects show that our pipeline preserves this advantage even under noisy real-world conditions.

**Keywords:** Inverse Rendering · 3D Reconstruction · Relighting

## 1 Introduction

Recent advances in radiance fields have achieved remarkable success in reconstructing scenes with high-fidelity appearance [28,36] and geometry [20,52]. However, these methods typically bake in the appearance under fixed lighting conditions, making them unsuitable for scenarios where relighting is required, e.g. placing reconstructed objects in virtual environments. While inverse rendering approaches estimate material and lighting separately for this purpose [16, 24], the decomposition achieved is often suboptimal due to the inherent ambiguity in inverse rendering. This ambiguity is traditionally resolved by capturing a static object under multiple lighting conditions [24, 32] or utilizing either handcrafted [16] or data-driven [7] priors. However, such methods require extra sophistication, from extended lighting representations for capturing varying lighting conditions [13, 24] to lengthy training procedures for learning useful priors from (limited) 3D data, either for surface materials [7], environment illumination [35] or relighting [40] directly. For most capture settings, a single far-field light representation can provide a decent approximation for the environment illumination. In this work, we posit that if an object is observed and reconstructed while undergoing significant motion, then the evolving light interaction between the object and a single far-field light representation provides enough observations to significantly enhance the material and lighting decomposition.

Since the single far-field light assumption requires tracking the object under large pose changes to observe significantly different light interactions, such dynamic reconstruction, while providing stronger constraints for material-lighting ambiguity, also demands accurate tracking, and tracking errors may degrade reconstruction quality. Although category-level or instance-level CAD models have been utilized [55] to track a rigidly-moving object in novel sequences, such methods assume that a reconstructed model is available and do not generalize to new objects, thus limiting their application. The most suitable setting for our purposes is the hand-held reconstruction setting, where a person freely rotates an object in their hands and a video is captured from a static viewpoint. Such a setting provides full 360 coverage of an object while also providing the varying lighting conditions we need to improve the material-lighting decomposition.

In this paper, we propose a relightable reconstruction pipeline that allows us to utilize the surface-light interactions of general dynamic objects for improved material estimation. We employ a three-stage approach to ease the optimization challenges of joint object tracking, reconstruction and inverse rendering. In the first stage, we use progressive sequential optimization together with the NeuS representation [52] to obtain the coarse geometry and initial poses of the dynamic object. We utilize 2D correspondences from state-of-the-art feature matchers [10, 30] to aid with this optimization. In the second stage, we introduce 3D Gaussians and optimize them jointly with the object poses to obtain fine-grained geometry and pose estimates. Given the reconstructed geometry, we augment the Gaussians with material properties in the final stage and employ physically-based rendering to optimize these materials across timesteps, enabling them to explain the observed surface appearance under changing illumination.

We further leverage efficient Gaussian ray tracing [37] to model accurate occlusion for the dynamic object within our inverse rendering pipeline. Finally, to quantitatively evaluate the improvement in material decomposition enabled by our 4D inverse rendering pipeline, we introduce a synthetic dataset of common household objects [2], along with ground-truth materials and relit views. We also demonstrate the practical viability of our method qualitatively on real-world hand-held object sequences.

Overall, we make the following contributions:

- We posit and empirically validate that observing an object under significant motion provides stronger constraints for material-lighting decomposition than a static capture.
- We propose a 4D inverse rendering framework that brings together object pose tracking and surface material estimation.
- We introduce a synthetic dataset of common household objects with ground-truth materials and relit views, enabling controlled evaluation of material-lighting decomposition quality in three common capture settings: static multiview, turntable and hand-held capture.

## 2 Related Work

Since our work brings together inverse rendering and dynamic object reconstruction, we briefly review state-of-the-art representations in the respective fields while also placing our work in the landscape of methods that specifically target better material-illumination decomposition.

**3D Reconstruction and Inverse Rendering.** Recent progress in inverse rendering has been tightly coupled with advances in differentiable 3D scene representations. NeRF [36] and 3D Gaussian Splatting [28]-based representations have become the de facto standard for high-quality novel view synthesis and 3D reconstruction. Since they bake view-dependent appearance into the geometry in their original formulation, methods [3, 16, 33, 62] quickly extended these representations with material properties and utilized physically-based rendering [26] to optimize these materials and environment illumination separately, allowing relighting. Improvements in the underlying representation, such as enhancing the underlying geometric fidelity of the reconstruction [20,52,58,59] and improving efficiency with hybrid data structures [6,38], were adopted by the inverse rendering methods [11,24,31,44,64], boosting their efficiency and accuracy. In conjunction with the representation, enhancements have been introduced in the physically-based inverse rendering pipeline, including more accurate sampling strategies [18, 19], varied shading functions [9, 23, 34], better illumination modelling [17, 56, 57, 61] and handling unconstrained image collections [4, 29]. The discrete nature of 3D-GS-based representations and the fast rendering have made them especially adept at capturing fine details and modelling computationally expensive global illumination effects—enabled by ray tracing of 3D Gaussians [37]—efficiently, both of which are crucial for accurate inverse rendering. However, in terms of modelling general objects, these representations have only

been used in the static setting so far. We extend them to the dynamic reconstruction setting.

**Material-Lighting Ambiguity.** A central challenge for inverse rendering methods is to resolve the ambiguity between surface reflectance and illumination. Under a fixed-illumination setup and without proper constraints, some of the object colour tends to get baked into the environment map, appearing as tints and shifts in tone and intensity. Vision science research has shown that humans are also bad at disambiguating between material and environment illumination [41]. Methods usually try to resolve this ambiguity through hand-crafted [16, 19] or learned [7, 35] priors, or capturing the object in a multi-illumination setting. The former strategy focuses on learning priors for either the surface material [7], environment illumination [35] or relighting directly [1, 63], hoping that the learned prior from large-scale datasets will constrain the optimization enough for proper disambiguation. The latter uses augmentation strategies, such as flashlight-enabled extra views [13], multi-environment captures [32] or modelling external occluders [48] to introduce additional lighting constraints to resolve ambiguity. Powerful, diffusion-based priors have also been utilized [40,46] to produce multi-light views from a single lighting condition, which are then utilized in the optimization. Instead of training on massive amounts of data or modelling complex lighting setups, our approach explores the orthogonal setting of dynamic object reconstruction, which provides abundant lighting conditions through a single RGB video and requires a simple far-field light representation to capture the gain in material-lighting decomposition.

**Hand-Held Object Reconstruction.** Inverse rendering methods usually require camera poses as input and assume a static scene. For dynamic, rigid reconstruction, the object has an additional six degrees of freedom to move in the observed scene. Most object pose tracking methods assume that an object model is available, either category-level [14, 47, 50, 51] or instance-level [53, 55], which allows them to track the object through relative pose estimation but limits their generalizability. For novel object pose tracking, methods usually rely on RGB-D input, combined with a NeRF [54] or Gaussian Splatting [25], to reconstruct and track the object. Some methods specialize in hand-held object tracking [8, 12], relying on hand pose annotations and consistent hand-object interactions to track the object. OnlineSplatter [21] introduces transformer-based 3D Gaussian tracking for real-time object-centric online reconstruction. More closely aligned with our setting of reconstructing and tracking free-moving objects in a globally consistent manner, the recent method FMOV [43] relies on NeuS and progressive pose optimization for the task, which is enhanced by utilizing correspondences from LoFTR [45], followed by a global refinement step. In contrast, our method uses the state-of-the-art matchers [10,30] for more reliable feature matching and introduces 3D-GS for the global refinement step, which allows fine-grained pose refinement since 3D Gaussians are better at capturing details, resulting in greater robustness for the inverse rendering task.

## 3 Motivation and Problem Setup

Our central thesis is that observing a dynamically moving object from a stationary camera yields a richer set of surface-light interactions than a moving camera capturing a static scene. Greater lighting diversity allows us to better condition the ill-posed inverse rendering problem, leading to better material estimates.

A common paradigm for capturing dynamic objects is the hand-held setup, where the user freely rotates a rigid object in front of a stationary camera. Let $\{I_t\}_{t=1}^{T}$ be a sequence of $T$ images observing a rigidly moving object with unknown per-frame poses $\{(R_t, t_t)\}_{t=1}^{T}$, with $(R_t, t_t) \in SE(3)$. For an observed image at time $t$, points $x$ on the canonical surface $S$ and their corresponding normals $\hat{n}$ are transformed according to the object pose as:

$$x_t = R_t x + t_t; \;\; \hat{n}_t = \hat{n}(x_t) = R_t \hat{n} \tag{1}$$

We assume that the environment does not change over time and—given the far-field illumination assumption—the environment light reaching a surface point $x$ on the object from direction $\hat{\omega}$ is given by:

$$L_i(x, \hat{\omega}) = L_{direct}(\hat{\omega}) V(x, \hat{\omega}) + L_{indirect}(x, \hat{\omega}) \tag{2}$$

where $V \in [0, 1]$ represents the visibility of the environment from the given surface point $x$ in direction $\hat{\omega}$, $L_{direct}$ is the far-field illumination and $L_{indirect}$ captures the global illumination effects. Combined with the spatially varying BRDF $f$, which defines the material at surface point $x$, the outgoing radiance at that point for a view direction $\hat{\omega}_o$ at time $t$ is defined as:

$$L_o(x, \hat{\omega}_o, t) = \int_{\Omega_t \subseteq \mathbb{S}^2} f(x_t, \hat{\omega}_{it}, \hat{\omega}_o) L_i(x_t, \hat{\omega}_{it}) (\hat{\omega}_{it} \cdot \hat{n}_t)_+ d\hat{\omega}_{it} \tag{3}$$

where $\Omega_t$ is the spherical domain of integration and $()_+$ is the positive clamping function.

Given a monocular video of a rigidly moving object, our goal is to recover the scene parameters, namely the object geometry $\{x, \hat{n}\} \in S$, material properties, i.e. albedo $a$ and roughness $r$, per-frame object poses $\{R_t, t_t\}_{t=1}^{T}$ and the environment illumination $L$. This is achieved by optimizing the following objective for each rendered pixel $u$ at time $t$:

$$\underset{x, \hat{n}, \{R,t\}_t, a, r, L}{\operatorname{argmin}} \sum_{t,u} |I_{\text{rend}}(u; x, \hat{n}, R_t, t_t, a, r, L) - I_t(u)| \tag{4}$$

where $I_{\text{rend}}(u, t)$ is rendered using a differentiable renderer and $I_t$ is the observed ground-truth image at time $t$.

From Equation 3, we can see that the canonical object surface points $x$ interact with the environment illumination $L$ across observations through the rotated surface normals $R_t \hat{n}$. Given high-frequency lighting and observations of the object undergoing significant pose changes, the same surface point is seen under numerous lighting conditions, introducing more constraints to the optimization in Equation 4 than the static-object multiview capture setting, where no such interactions are observed.

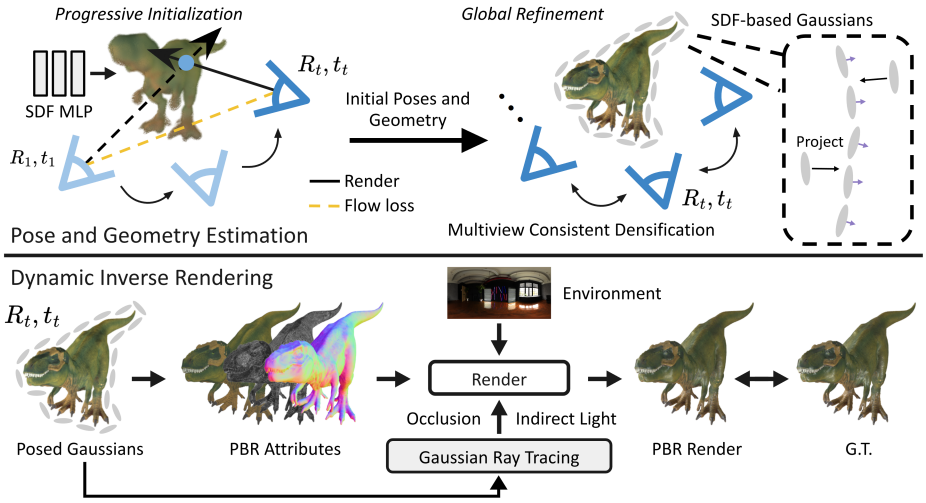


**Fig. 2:** Overview of our three-stage method. We first perform progressive online optimization with a NeuS representation to estimate initial object poses and geometry. We then refine them with an SDF-based 3D Gaussian representation to capture fine-grained geometry. Finally, we add surface materials to the Gaussians and jointly optimize them with environment illumination to obtain accurate material-lighting decomposition from a moving object.

## 4 Method

The objective in Equation 4 is severely ill-posed and estimating all unknown scene parameters jointly from the start is prone to getting stuck in suboptimal local minima. To tackle this problem, we introduce a multi-stage pipeline, as shown in Figure 2. First, we obtain the initial geometry and object poses through a progressive optimization scheme that uses a neural SDF representation for the object (Section 4.1). Next, we switch to the 3D Gaussian representation and refine the initial estimates jointly across all frames to capture the fine details (Section 4.2). Finally (Section 4.3), we add materials to our representation and model the environment illumination, optimizing them jointly with the geometry and poses through physically-based rendering to accurately decompose material and lighting from observations of a moving object.

### 4.1 Progressive Pose and Geometry Estimation

Estimating the pose and geometry is a classic ill-posed problem, and we need to optimize both iteratively to achieve robust solutions. We take inspiration from FMOV [43] to estimate object poses progressively with a neural SDF representation. The key insight here is to learn the geometry on a coarse level, to prevent the poses from overfitting to high-frequency noise, while still getting a sufficiently good initial estimate for our second-stage refinement.

**Neural SDF Representation.** To facilitate progressive pose optimization, we adopt NeuS [43, 52] to initially represent the geometry, making use of the regularization properties of MLPs and the unit sphere geometric initialization of the SDF network, which keeps the surface intact under sparse observations. Given a point sample $x$ along the view direction $v$, the object's surface and appearance are modelled as:

$$s(x), c(x, v) = F(x, v; \theta) \tag{5}$$

where $F$ is an MLP network with parameters $\theta$, $c$ is the view-dependent colour and $s$ is the SDF value at point $x$. The rendered colour for a pixel $u$ is derived using the standard quadrature approximation of the volume rendering integral [36]:

$$I(u) = \sum_{k=1}^{K} w_k c_k \tag{6}$$

where $K$ is the number of samples along the ray—kept low here to model coarse geometry—and colour $c_k$ is alpha-blended using blending weights $w_k$, composed of the transmittance and the volume density derived from the underlying SDF value [52].

**Per-Frame Pose Representation.** Since the camera/object pose across the captured image sequence is unknown, we jointly optimize it with the geometry. The camera rays used to query the SDF MLP are transformed to each timestep $t$ using the pose $(R_t, t_t)$ predicted by a per-frame pose MLP. We also adopt the virtual camera system from FMOV [43], which eliminates two degrees of freedom in translation by centering the object in each image using 2D masks. Although not physically accurate, this eases the pose optimization for noisy, real-world sequences.

**Sequential Optimization.** We sequentially optimize the geometry and track the pose frame-by-frame—where each frame is initialized by the pose of the previous frame—to handle large pose changes and also reset the geometry once the pose has changed significantly, so that the degraded surface quality does not hinder further pose tracking [43]. We also use 2D feature correspondences [10,30] in the neighbourhood of the current frame to constrain tracking by imposing cross-frame consistency; given that hand-held objects are often textureless and specular, the prior from a learned matcher helps regularize the tracking in such cases. The final loss at this stage is:

$$\mathcal{L}_{progressive} = \mathcal{L}_{render} + \mathcal{L}_{eikonal} + \mathcal{L}_{mask} + \mathcal{L}_{match} \tag{7}$$

where $\mathcal{L}_{render}$, $\mathcal{L}_{eikonal}$ and $\mathcal{L}_{mask}$ are the photometric, eikonal and mask losses from NeuS [52] and $\mathcal{L}_{match}$ is the flow loss from FMOV [43]. At each step, only the pose MLPs of the current frame and the frame with matched correspondences are optimized.

### 4.2 Global Pose and Geometry Refinement

Given the progressively optimized poses from the previous stage, we can now jointly refine them with the geometry at a finer level to obtain a more detailed

reconstruction. Surface normals are also estimated for the geometry at this stage, and obtaining a good estimate is crucial for accurate material estimation later.

**3D Gaussian Representation.** For the refinement stage, we introduce 3D Gaussians [28] as our geometry representation since they are good at capturing fine details. We integrate an SDF directly with the Gaussians, as introduced in DiscretizedSDF [64], which helps stabilize the geometry during inverse rendering, especially for shinier objects. Each Gaussian is defined as an ellipsoid with a position $\mu$, a covariance matrix $\Sigma$ (parameterized as a quaternion rotation $q$ and a 3D scaling factor $s$), an SDF value $\sigma$ and view-dependent colour $c$ represented using spherical harmonics. The normal of a Gaussian is defined as the rotation axis corresponding to the smallest scale. The opacity is derived from the SDF as:

$$O(\sigma) = \frac{4e^{-\gamma\sigma}}{(1+e^{-\gamma\sigma})^2} \tag{8}$$

with learnable global variance $\gamma$, while the response of the Gaussian $G$ is defined as:

$$G(x) = \exp(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)) \tag{9}$$

During rendering, 3D Gaussians are projected to the 2D image plane and alpha composited using Equation 6, with the blending weights $w_k$ determined by the SDF-derived opacity $O(\sigma)$ and the response $G(u)$ for the projected 2D Gaussian at pixel $u$. Gaussians with backfacing normals are culled during rasterization, which avoids entanglement of front and back surfaces—always well-defined through the use of sphere initialization [64]—and helps solidify the SDF on the correct surface only [44]. We also use geometry-aware rasterization [60] for rendering depth, which is determined directly by the intersection point of a ray with a Gaussian in 3D instead of using the Gaussian center, thus improving fidelity.

To render a Gaussian at time $t$, we first transform it using the object pose $(R_t, t_t)$. This results in transforming the position $\mu$ and rotation $q$ as:

$$\mu_t = R_t\mu + t_t; \;\; q_t = Q(R_t)q \tag{10}$$

where $Q(\cdot)$ converts a rotation matrix to a quaternion. Note that since the normal is derived from the Gaussian itself, it is already in the correct orientation when the transformed Gaussian is rendered. The spherical harmonics are evaluated equivalently by rotating the query direction with $R_t^T$ instead. Transforming the Gaussians rather than the cameras also allows the gradients to flow to the pose MLP for joint optimization [15].

**Global Pose Representation.** For pose refinement, we use a global MLP with per-pose embeddings to learn deviations from the poses estimated during the first stage. Since the estimates from the previous stage were obtained using virtual cameras, an initialization for the full 6-DoF poses is computed from them using 3D-2D correspondences from the estimated mesh with the RANSAC EPnP algorithm [43].

**Adaptive Density Control.** We incorporate the multi-view consistent densification and pruning strategy introduced by FastGS [42] into our pipeline. This keeps the number of Gaussians low and consolidates them at the correct locations, which improves the efficiency of the compute-intensive inverse rendering and 3D Gaussian ray tracing in the next stage.
**Optimization.** At this stage, we globally optimize the geometry and all object poses using the following objective:

$$\mathcal{L}_{global} = \mathcal{L}_{render} + \mathcal{L}_{normal} + \mathcal{L}_{variance} + \mathcal{L}_{rank} + \mathcal{L}_{\gamma} + \mathcal{L}_{proj} + \mathcal{L}_{mask} \quad (11)$$

where $\mathcal{L}_{render}$ is the photometric loss, $\mathcal{L}_{normal}$ is the regularization on rendered normals [20], $\mathcal{L}_{variance}$ forces the rendered depth variance to be small [16], $\mathcal{L}_{rank}$ forces the Gaussians to be disc-like [22], further improving their geometric fidelity, and $\mathcal{L}_{\gamma}$ and $\mathcal{L}_{proj}$ are the median and projection losses for the SDF [64].

### 4.3 Dynamic Inverse Rendering

Given a reliable estimate of the geometry from the second stage, we introduce a material model on top of the 3D Gaussian representation and jointly optimize it with the environment illumination to recover accurate, disentangled materials from observations of the object under varying poses.
**Material and Environment Representation.** We parameterize the spatially varying BRDF $f$ with a diffuse and a specular term. The diffuse term is modelled with a Lambertian reflectance model as $f_d = \frac{a}{\pi}$, where $a$ is the albedo, and the specular term is represented by the GGX microfacet model [49] as:

$$f_s(x_t, \hat{\omega}_{it}, \hat{\omega}_o, r, F_0) = \frac{DFG}{4 \cdot (\hat{\omega}_i \cdot \hat{n}) \cdot (\hat{\omega}_o \cdot \hat{n})} \quad (12)$$

where $r$ is the roughness, $F_0$ is the specular reflectance at normal incidence and $D$, $F$ and $G$ are the normal distribution, Fresnel and geometry attenuation functions. The final BRDF is represented as $f = f_d + f_s$. We add the optimizable material parameters $a \in \mathbb{R}^3$ and $r \in \mathbb{R}$ to the Gaussians at this stage, which can be rendered into material maps similar to the other Gaussian properties. We assume dielectric materials with $F_0 = 0.04$. The material is made spatially varying by storing four values for each material property and the normal at the corners of the Gaussian plane, which are then interpolated at the intersection point, as introduced in SVG-IR [44], thus improving representation capacity and accuracy. We assume that the illumination is distant and can be represented as an environment map. Since a high resolution environment map can be used as an error sink early on in the optimization, we start at a low resolution and bilinearly upsample it progressively to a higher resolution.
**Visibility and Global Illumination.** Since we want to consolidate multiple surface-light interactions into consistent material estimates, it is important to model self-occlusions for an object. We use the ray tracer for Gaussians introduced by 3DGRT [37] to model such occlusions in our pipeline. Given secondary

rays originating from surface point $x$ towards sampled incident light directions $\hat{\omega}_i$, the tracer is queried as

$$(c_i, o_i) \leftarrow Trace(x, \hat{\omega}_i) \tag{13}$$

where the visibility $V$ can be deduced by thresholding the rendered opacity $o_i$ and the rendered colour $c_i$ is used as an approximation of the indirect illumination coming from $\omega_i$. Plugging these values into Equation 2 gives us the final incoming illumination.

**Rendering and Optimization.** We adopt the deferred shading approach in our pipeline. We first render the Gaussians, transformed with object pose $(R_t, t_t)$ for timestep $t$, to get the depth, normal and material maps. These maps are then used to perform physically-based rendering at the estimated surface point using Equation 3. We use Fibonacci-based sphere sampling for rendering which adds bias of a fixed pattern but provides optimization stability [57]. The final loss is defined as:

$$\mathcal{L}_{pbr} = \mathcal{L}_{render} + \mathcal{L}_{light} + \mathcal{L}_{material} \tag{14}$$

where $\mathcal{L}_{light}$ and $\mathcal{L}_{material}$ define the smoothness losses on the environment and material maps [16].

**Relighting.** Since optimization instability is no longer a concern during relighting, we use BRDF- and lighting-based importance sampling [39] to get high-quality relights. As the optimized radiance of the Gaussians becomes invalid for indirect illumination during relighting, we use the split-sum approximation [27] to efficiently compute one-bounce indirect radiance for improved fidelity of relights [17].

## 5 Experiments

In the following, we describe our experimental setup, extensively evaluate the material-lighting decomposition performance and benchmark our method against existing baselines. See the supplementary material for implementation details.

**Synthetic Data.** HOT3D [2] provides a commonly used dataset of hand-held object interactions. As evaluating our method requires ground-truth material properties and relit views, we use the object assets from this dataset to generate our own renders with known ground-truth materials and illumination. Based on these assets, we create the following settings designed to mimic common real-world capture scenarios:

- **Static Dome**  **:** The object remains static while the camera moves around it, capturing views from the upper hemisphere.
- **Turntable**  **:** This setting models the classical turntable scenario, where the camera is static and the object rotates around a single axis. While this introduces varying illumination conditions, the motion is limited to a single degree of freedom.

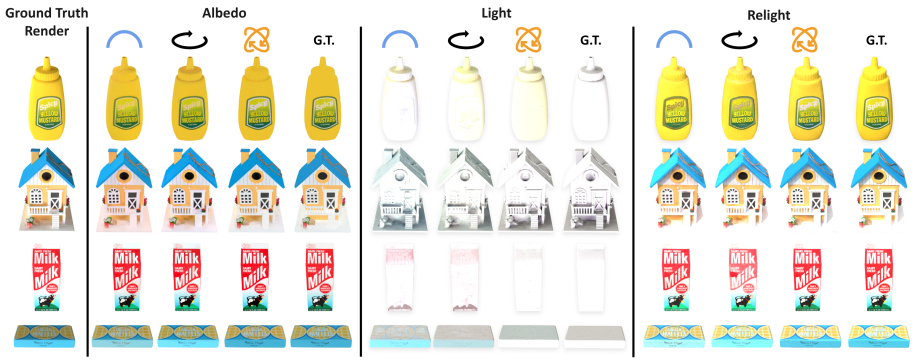


**Fig. 3:** Qualitative results for different capture configurations on our proposed synthetic dataset for the diffuse variant. The improved disentanglement of albedo and lighting (note the progression of disentanglement from left to right) induced by the hand-held setting results in better relighting consistently across a diverse set of objects.

– **Hand-held Rotations** : This is our proposed setting, where the user holds the object in front of the camera and rotates it around arbitrary axes. In the synthetic data, we do not explicitly model the hand, but instead render the object undergoing arbitrary rotations.

We select ten objects with varying geometries and render two material variants: an original variant using the material textures provided with the HOT3D assets, which are generally more specular, and a diffuse variant that keeps the albedo texture but uses a Lambertian material model, where material and lighting are harder to disentangle. The camera trajectories for the three capture settings are designed to share a common set of views, which we use for evaluation.

**Real Data.** For a valid comparison between capture settings, we require the object to be captured under the same lighting conditions for each setting. Since none of the existing real-world benchmarks satisfy this requirement, we record two real hand-held objects in the static dome and hand-held settings under the same environment lighting. We use SAM3 [5] to obtain masks for the captured objects.

**Metrics.** We use PSNR, SSIM, and LPIPS for albedo and relighting to assess decomposition quality, mean angular error (MAE) for normals and RMSE for environment maps. We normalize the albedo and relighting scale using ground-truth albedo for synthetic evaluation [62].

### 5.1 Material-Lighting Decomposition Evaluation

**Synthetic Data.** Figure 3 provides a qualitative comparison for the three common capture trajectories reproduced in our synthetic dataset. In the static dome setting, the fixed incoming illumination observed at each surface point leads to

**Table 1:** Quantitative results for different capture configurations on our proposed dataset. Results are reported for original and diffuse object variants. Despite evaluation with ground-truth poses, the static setting struggles to disentangle material and lighting, especially for the diffuse variant, while the hand-held setting performs well on both variants and with both estimated and ground-truth poses. **Bold** indicates the best result, and underline indicates the second-best result.

| Setting | Pose | Albedo | | | Relighting | | | Normal |
|---|---|---|---|---|---|---|---|---|
| | | PSNR↑ | SSIM↑ | LPIPS↓ | PSNR↑ | SSIM↑ | LPIPS↓ | MAE↓ |
| Original | | | | | | | | |
|  | G.T. | 36.72 | 0.987 | 0.018 | 38.14 | 0.986 | 0.020 | 0.83 |
|  | Est. | 34.10 | 0.983 | 0.021 | 33.02 | 0.977 | 0.027 | 2.45 |
| | G.T. | 36.15 | 0.987 | 0.016 | 36.01 | 0.983 | 0.022 | 1.20 |
|  | Est. | 40.37 | **0.991** | **0.013** | 38.54 | 0.988 | 0.017 | 1.56 |
| | G.T. | **40.72** | **0.991** | **0.013** | **40.24** | **0.989** | **0.016** | **0.53** |
| Diffuse | | | | | | | | |
|  | G.T. | 32.68 | 0.983 | 0.021 | 34.85 | 0.981 | 0.024 | 1.59 |
|  | Est. | 32.97 | 0.982 | 0.022 | 33.04 | 0.976 | 0.027 | 1.83 |
| | G.T. | 35.65 | 0.984 | 0.016 | 35.10 | 0.981 | 0.023 | 1.42 |
|  | Est. | 40.33 | **0.990** | **0.010** | 40.37 | 0.990 | **0.015** | 0.75 |
| | G.T. | **40.55** | **0.990** | **0.010** | **41.29** | **0.991** | **0.015** | **0.58** |

a stronger ambiguity, causing the albedo colour to bleed into the illumination. The turntable case increases the surface-light interactions for the surface points but they are still comparatively limited since the rotation is constrained to a single axis. The hand-held rotation setting, in contrast, exposes each surface point to a diverse set of light interactions, which imposes stronger constraints on the proper separation of albedo and lighting, and yields much better renders under novel illumination.

The quantitative results are provided in Table 1. We give full advantage of ground-truth camera poses to the static setting and evaluate the turntable and hand-held settings both with ground-truth and estimated poses. Notably, the hand-held setting with estimated poses outperforms the static setting with ground-truth poses despite having worse normals, highlighting the advantage of using the hand-held setting. In general, the hand-held setting achieves the strongest decomposition performance and the best relighting metrics, supporting our hypothesis that motion helps disentangle material and illumination during inverse rendering. The turntable setting additionally suffers from unconstrained normal reconstruction due to limited view coverage.

**Real Data.** We demonstrate the advantage of the hand-held setting over the common static casual capture setting on two real captured objects in Figure 4.

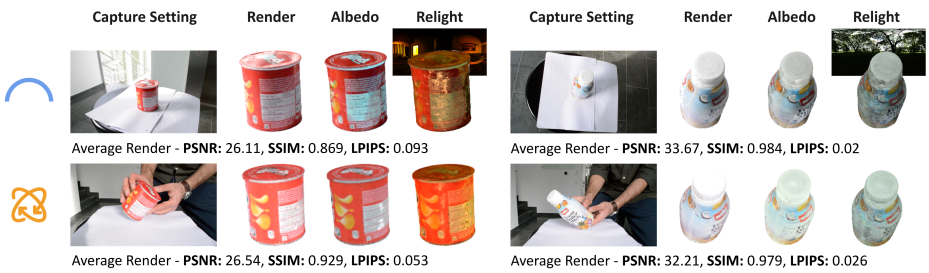


**Fig. 4:** Real captures under static multiview (top) and hand-held (bottom) settings. Despite similar render quality, hand-held capture yields cleaner albedo and more plausible relighting.

Despite similar rendering quality across the full sequences, the decomposition is qualitatively better in the hand-held setting. The static setting bakes shadows from the render into the albedo and tints the white regions, while the hand-held setting produces flatter and more consistent albedo, as expected, which in turn results in more plausible relights. This shows the viability of our method in real-world scenarios.

**Material Roughness Analysis.** To analyze how the material properties affect decomposition ambiguity, we create different versions of one object in our synthetic dataset by varying the roughness while keeping the albedo fixed. The resulting objects are rendered under the same environment map using the static and hand-held settings. Figure 5 shows the reconstructed environment maps from the resulting sequences, which most clearly illustrates the effect of different roughness levels. Lower roughness already helps disambiguation in the static setting since the sharper highlights provide stronger constraints and cannot be explained by the diffuse albedo alone. As the roughness increases and the reflections become more diffuse, it becomes harder to recover the illumination signal, increasing the tendency of the static setting to trade off albedo and illumination. The diverse surface-light interactions in the hand-held setting provide stronger directionality—which also curbs the swapping tendency of albedo and illumination—and enable the model to recover more information even in the challenging diffuse case while retaining the benefit across the roughness range.

**Illumination Analysis.** We further study the effect of different environment maps on the material decomposition quality for the static and hand-held settings. Figure 6 illustrates this by reconstructing the object from views rendered with different environment maps. In the case of uniform white light, no additional surface-light interactions are available to the hand-held setting, while the estimated albedo from the static setting degrades only mildly due to the reduced lighting complexity. As we progress to environments with more varying lighting, the albedo disentanglement starts degrading in the static setting due to increased lighting complexity while the hand-held setting, due to the stronger constraints from increasingly diverse surface-light interactions, is able to maintain the albedo quality.

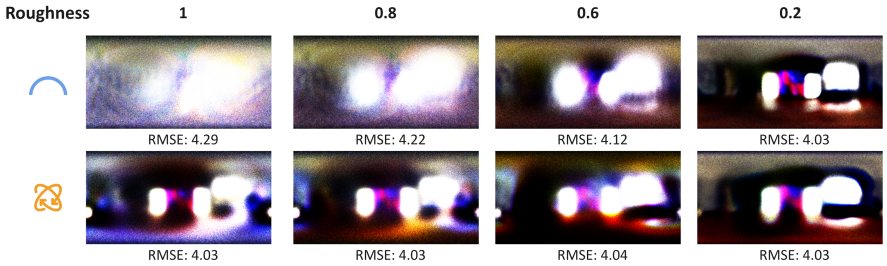


**Fig. 5:** Environment reconstruction under varying roughness. The hand-held setting recovers sharper illumination structure even for diffuse objects, while the gap to the static setting narrows for more specular objects, where sharper highlights provide stronger lighting constraints. The ground truth environment map is shown in Figure 2.

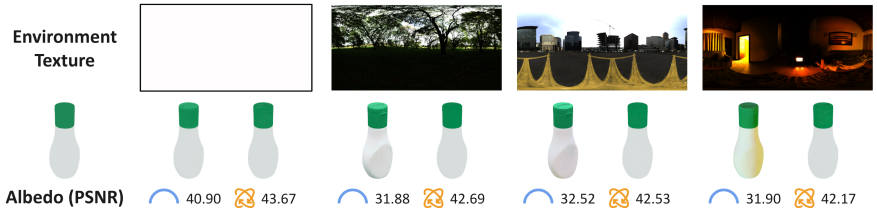


**Fig. 6:** Effect of illumination complexity on decomposition. The hand-held setting maintains high albedo quality under increasingly complex lighting, while the static setting degrades.

### 5.2 Benchmarking Relighting Performance

In our work, we introduce the new hand-held setting that requires object tracking. We validate our design choices in two ways: (1) performance in the static setting on a standard benchmark and (2) performance in the dynamic setting against adapted baselines. See the supplementary material for our pose estimation performance.

**Static Setting.** Table 2 shows the evaluation of albedo reconstruction, relighting under novel illumination and normal accuracy on the standard TensoIR benchmark [24]. Our method is competitive with baselines across metrics, achieving improved albedo and relighting quality. This suggests that our pipeline is already capable of achieving the strongest decomposition results for static inverse rendering, and motivates using the same pipeline in the dynamic setting where additional motion-induced lighting constraints further help improve the decomposition.

**Dynamic Setting.** As there are no existing baselines for test-time dynamic inverse rendering of general objects, we adapt existing methods for comparison on our dataset in Table 3. ReCap [32] optimizes a separate environment map for each capture illumination. We treat each frame of a dynamic scene as a separate lighting condition and optimize a separate environment for it (shared over consecutive frames for efficiency). Such a setup is not suitable for dynamic scenes:

**Table 2:** Decomposition results on the synthetic TensoIR dataset, where we outperform the static-setting baselines.

| Method | Albedo PSNR ↑ | Relighting PSNR ↑ | Normal MAE ↓ |
|---|---|---|---|
| SVG-IR | 30.27 | 31.19 | 4.19 |
| IRGS | 33.41 | 30.63 | **3.99** |
| Ours | **33.96** | **32.17** | 4.19 |

**Table 3:** Decomposition results on four hand-held sequences of our dataset. *Extended to handle rotating objects.

| Method | Envs | Albedo PSNR↑ | Relighting PSNR↑ | Normal MAE↓ |
|---|---|---|---|---|
| ReCap | 10 | 21.91 | 25.74 | 8.06 |
| | 40 | 22.42 | 25.58 | 8.01 |
| | 100 | 22.31 | 25.20 | 8.08 |
| SVG-IR* | 1 | 33.40 | 32.22 | 3.69 |
| IRGS* | 1 | 40.26 | 32.99 | 0.56 |
| Ours | 1 | **41.03** | **39.65** | **0.51** |

sharing more frames per environment approximates too much while less sharing runs into optimization issues. Designed for static scenes, we extend SVG-IR [44] and IRGS [17] to account for object dynamics when querying the environment map. Our consolidated SDF-based geometry modelling yields the strongest decomposition and subsequent relighting performance in the hand-held setting.

## 6 Conclusion

Instead of introducing additional light sources or relying on strong learned priors, we investigate the long-standing material-lighting ambiguity in inverse rendering from the perspective of leveraging the natural diversity of surface-light interactions induced by rigid object motion. Our proposed relightable 4D reconstruction pipeline leverages these interactions to improve decomposition in the challenging hand-held capture setting. Across controlled experiments on our proposed synthetic dataset, we show that increasing object pose diversity yields stronger constraints for disentangling albedo and illumination, while we also demonstrate the feasibility of our method in improving decomposition for real hand-held objects.

**Future Work.** Modelling hand occlusions and more complex near-field light interactions in real scenes is an exciting future direction. Moreover, non-rigid objects can provide even richer motion diversity, making a general non-rigid inverse rendering pipeline a natural next step.

## Acknowledgements

Jan Eric Lenssen is supported by the German Research Foundation (DFG) - 556415750 (Emmy Noether Programme, project: Spatial Modeling and Reasoning).

## References


1. Alzayer, H., Henzler, P., Barron, J.T., Huang, J.B., Srinivasan, P.P., Verbin, D.: Generative multiview relighting for 3d reconstruction under extreme illumination variation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10933–10942 (June 2025)
2. Banerjee, P., Shkodrani, S., Moulon, P., Hampali, S., Han, S., Zhang, F., Zhang, L., Fountain, J., Miller, E., Basol, S., Newcombe, R., Wang, R., Engel, J.J., Hodan, T.: Hot3d: Hand and object tracking in 3d from egocentric multi-view videos. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 7061–7071 (June 2025)
3. Boss, M., Braun, R., Jampani, V., Barron, J.T., Liu, C., Lensch, H.: Nerd: Neural reflectance decomposition from image collections. In: IEEE/CVF International Conference on Computer Vision (ICCV). pp. 12684–12694 (2021)
4. Boss, M., Engelhardt, A., Kar, A., Li, Y., Sun, D., Barron, J., Lensch, H., Jampani, V.: Samurai: Shape and material from unconstrained real-world arbitrary image collections. Advances in Neural Information Processing Systems (NeurIPS) **35**, 26389–26403 (2022)
5. Carion, N., Gustafson, L., Hu, Y.T., Debnath, S., Hu, R., Coll-Vinent, D.S., Ryali, C., Alwala, K.V., Khedr, H., Huang, A., Lei, J., Ma, T., Guo, B., Kalla, A., Marks, M., Greer, J., Wang, M., Sun, P., Rädle, R., Afouras, T., Mavroudi, E., Xu, K., Wu, T.H., Zhou, Y., Momeni, L., HAZRA, R., Ding, S., Vaze, S., Porcher, F., Li, F., Li, S., Kamath, A., Cheng, H.K., Dollar, P., Ravi, N., Saenko, K., Zhang, P., Feichtenhofer, C.: SAM 3: Segment anything with concepts. In: International Conference on Learning Representations (ICLR) (2026), https://openreview.net/forum?id=r35clVtGzw
6. Chen, A., Xu, Z., Geiger, A., Yu, J., Su, H.: Tensorf: Tensorial radiance fields. In: European Conference on Computer Vision (ECCV) (2022)
7. Chen, X., Peng, S., Yang, D., Liu, Y., Pan, B., Lv, C., Zhou, X.: Intrinsicanything: Learning diffusion priors for inverse rendering under unknown illumination. In: European Conference on Computer Vision (ECCV). p. 450–467. Springer-Verlag, Berlin, Heidelberg (2024). https://doi.org/10.1007/978-3-031-73027-6_26, https://doi.org/10.1007/978-3-031-73027-6_26
8. Chen, Z., Potamias, R.A., Chen, S., Schmid, C.: HORT: Monocular hand-held objects reconstruction with transformers. In: IEEE/CVF International Conference on Computer Vision (ICCV) (2025)
9. Dihlmann, J.N., Majumdar, A., Engelhardt, A., Braun, R., Lensch, H.: Subsurface scattering for gaussian splatting. Advances in Neural Information Processing Systems (NeurIPS) **37**, 121765–121789 (2024)
10. Edstedt, J., Nordström, D., Zhang, Y., Bökman, G., Astermark, J., Larsson, V., Heyden, A., Kahl, F., Wadenbäck, M., Felsberg, M.: Roma v2: Harder better faster denser feature matching (2025), https://arxiv.org/abs/2511.15706
11. Engelhardt, A., Raj, A., Boss, M., Zhang, Y., Kar, A., Li, Y., Sun, D., Brualla, R.M., Barron, J.T., Lensch, H., et al.: Shinobi: Shape and illumination using neural object decomposition via brdf optimization in-the-wild. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 19636–19646 (2024)
12. Fan, Z., Parelli, M., Kadoglou, M.E., Chen, X., Kocabas, M., Black, M.J., Hilliges, O.: Hold: Category-agnostic 3d reconstruction of interacting hands and objects from video. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 494–504 (2024)

13. Fei, F., Tang, J., Tan, P., Shi, B.: Vminer: Versatile multi-view inverse rendering with near- and far-field light sources. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 11800–11809 (June 2024)
14. Fischer, T., Zhang, X., Ilg, E.: Unified category-level object detection and pose estimation from rgb images using 3d prototypes. In: IEEE/CVF International Conference on Computer Vision (ICCV). pp. 9790–9800 (2025)
15. Fu, Y., Wang, X., Liu, S., Kulkarni, A., Kautz, J., Efros, A.A.: Colmap-free 3d gaussian splatting. In: CVPR. pp. 20796–20805. IEEE (2024)
16. Gao, J., Gu, C., Lin, Y., Li, Z., Zhu, H., Cao, X., Zhang, L., Yao, Y.: Relightable 3d gaussians: Realistic point cloud relighting with brdf decomposition and ray tracing. In: Leonardis, A., Ricci, E., Roth, S., Russakovsky, O., Sattler, T., Varol, G. (eds.) European Conference on Computer Vision (ECCV). pp. 73–89. Springer Nature Switzerland, Cham (2025)
17. Gu, C., Wei, X., Zeng, Z., Yao, Y., Zhang, L.: Irgs: Inter-reflective gaussian splatting with 2d gaussian ray tracing. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10943–10952 (2025)
18. Gu, C., Wei, X., Zhang, L., Zhu, X.: Tensoflow: Tensorial flow-based sampler for inverse rendering. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 495–504 (2025)
19. Hasselgren, J., Hofmann, N., Munkberg, J.: Shape, light, and material decomposition from images using monte carlo rendering and denoising. In: Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., Oh, A. (eds.) Advances in Neural Information Processing Systems (NeurIPS). vol. 35, pp. 22856–22869. Curran Associates, Inc. (2022), https://proceedings.neurips.cc/paper_files/paper/2022/file/8fcb27984bf16ca03cad643244ec470d-Paper-Conference.pdf
20. Huang, B., Yu, Z., Chen, A., Geiger, A., Gao, S.: 2d gaussian splatting for geometrically accurate radiance fields. In: ACM SIGGRAPH. SIGGRAPH '24, Association for Computing Machinery, New York, NY, USA (2024). https://doi.org/10.1145/3641519.3657428, https://doi.org/10.1145/3641519.3657428
21. Huang, M.H., Foo, L.G., Theobalt, C., Sun, Y., Soh, D.W.: Onlinesplatter: Pose-free online 3d reconstruction for free-moving objects. In: Advances in Neural Information Processing Systems (NeurIPS) (2025)
22. Hyung, J., Hong, S., Hwang, S., Lee, J., Choo, J., Kim, J.: Effective rank analysis and regularization for enhanced 3d gaussian splatting. In: NeurIPS (2024)
23. Jiang, Y., Tu, J., Liu, Y., Gao, X., Long, X., Wang, W., Ma, Y.: Gaussianshader: 3d gaussian splatting with shading functions for reflective surfaces. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5322–5332 (2024)
24. 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 (June 2023)
25. Jin, Y., Prasad, V., Jauhri, S., Franzius, M., Chalvatzaki, G.: 6dope-gs: Online 6d object pose estimation using gaussian splatting. In: IEEE/CVF International Conference on Computer Vision (ICCV). pp. 8032–8043 (2025)
26. Kajiya, J.T.: The rendering equation. ACM SIGGRAPH Computer Graphics **20**(4), 143–150 (1986). https://doi.org/10.1145/15886.15902
27. Karis, B., Games, E.: Real shading in unreal engine 4. Proc. Physically Based Shading Theory Practice **4**(3), 1 (2013)
28. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G.: 3d gaussian splatting for real-time radiance field rendering. ACM Transactions on Graphics (ToG) **42**(4), 1–14 (2023)

29. Kuang, Z., Olszewski, K., Chai, M., Huang, Z., Achlioptas, P., Tulyakov, S.: Neroic: Neural rendering of objects from online image collections. ACM Transactions on Graphics (ToG) **41**(4), 1–12 (2022)
30. Leroy, V., Cabon, Y., Revaud, J.: Grounding image matching in 3d with mast3r. In: European Conference on Computer Vision (ECCV). p. 71–91. Springer-Verlag, Berlin, Heidelberg (2024). https://doi.org/10.1007/978-3-031-73220-1_5, https://doi.org/10.1007/978-3-031-73220-1_5
31. Li, J., Wang, L., Zhang, L., Wang, B.: Tensosdf: Roughness-aware tensorial representation for robust geometry and material reconstruction. ACM Transactions on Graphics (ToG) **43**(4) (Jul 2024). https://doi.org/10.1145/3658211, https://doi.org/10.1145/3658211
32. Li, J., Wu, Z., Zamfir, E., Timofte, R.: Recap: Better gaussian relighting with cross-environment captures. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 21307–21316 (June 2025)
33. Liang, Z., Zhang, Q., Feng, Y., Shan, Y., Jia, K.: Gs-ir: 3d gaussian splatting for inverse rendering. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 21644–21653 (2024)
34. Liu, Y., Wang, P., Lin, C., Long, X., Wang, J., Liu, L., Komura, T., Wang, W.: Nero: Neural geometry and brdf reconstruction of reflective objects from multiview images. ACM Transactions on Graphics (ToG) **42**(4), 1–22 (2023)
35. Lyu, L., Tewari, A., Habermann, M., Saito, S., Zollhöfer, M., Leimkühler, T., Theobalt, C.: Diffusion posterior illumination for ambiguity-aware inverse rendering. ACM Transactions on Graphics (ToG) **42**(6) (Dec 2023). https://doi.org/10.1145/3618357, https://doi.org/10.1145/3618357
36. 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) (2020)
37. Moenne-Loccoz, N., Mirzaei, A., Perel, O., de Lutio, R., Martinez Esturo, J., State, G., Fidler, S., Sharp, N., Gojcic, Z.: 3d gaussian ray tracing: Fast tracing of particle scenes. ACM Transactions on Graphics (ToG) **43**(6) (Nov 2024). https://doi.org/10.1145/3687934, https://doi.org/10.1145/3687934
38. Müller, T., Evans, A., Schied, C., Keller, A.: Instant neural graphics primitives with a multiresolution hash encoding. ACM Transactions on Graphics (ToG) **41**(4) (Jul 2022). https://doi.org/10.1145/3528223.3530127, https://doi.org/10.1145/3528223.3530127
39. Pharr, M., Jakob, W., Humphreys, G.: Physically based rendering: From theory to implementation. MIT Press (2023)
40. Poirier-Ginter, Y., Gauthier, A., Phillip, J., Lalonde, J.F., Drettakis, G.: A diffusion approach to radiance field relighting using multi-illumination synthesis. Computer Graphics Forum **43**(4), e15147 (2024). https://doi.org/https://doi.org/10.1111/cgf.15147, https://onlinelibrary.wiley.com/doi/abs/10.1111/cgf.15147
41. Pont, S.C., te Pas, S.F.: Material — illumination ambiguities and the perception of solid objects. Perception **35**(10), 1331–1350 (2006). https://doi.org/10.1068/p5440, https://doi.org/10.1068/p5440, pMID: 17214380
42. Ren, S., Wen, T., Fang, Y., Lu, B.: Fastgs: Training 3d gaussian splatting in 100 seconds. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 26094–26103 (2026)
43. Shi, H., Hu, Y., Koguciuk, D., Lin, J.T., Salzmann, M., Ferstl, D.: Free-moving object reconstruction and pose estimation with virtual camera. AAAI Conference on Artificial Intelligence **39**(7), 6860–6868 (Apr 2025). https://doi.org/

10.1609/aaai.v39i7.32736, https://ojs.aaai.org/index.php/AAAI/article/view/32736
44. Sun, H., Gao, Y., Xie, J., Yang, J., Wang, B.: Svg-ir: Spatially-varying gaussian splatting for inverse rendering. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 16143–16152 (2025)
45. Sun, J., Shen, Z., Wang, Y., Bao, H., Zhou, X.: Loftr: Detector-free local feature matching with transformers. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 8922–8931 (2021)
46. Tang, J., Levine, M., Verbin, D., Garbin, S.J., Niessner, M., Martin-Brualla, R., Srinivasan, P.P., Henzler, P.: ROGR: Relightable 3D Objects using Generative Relighting. In: Advances in Neural Information Processing Systems (NeurIPS) (2025)
47. Tian, X., Lin, X., Zhong, F., Qin, X.: Large-displacement 3d object tracking with hybrid non-local optimization. In: European Conference on Computer Vision (ECCV). pp. 627–643. Springer (2022)
48. Verbin, D., Mildenhall, B., Hedman, P., Barron, J.T., Zickler, T., Srinivasan, P.P.: Eclipse: Disambiguating illumination and materials using unintended shadows. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 77–86 (June 2024)
49. Walter, B., Marschner, S.R., Li, H., Torrance, K.E.: Microfacet models for refraction through rough surfaces. Rendering techniques **2007**, 18th (2007)
50. Wang, C., Martín-Martín, R., Xu, D., Lv, J., Lu, C., Fei-Fei, L., Savarese, S., Zhu, Y.: 6-pack: Category-level 6d pose tracker with anchor-based keypoints. In: IEEE International Conference on Robotics and Automation (ICRA). pp. 10059–10066. IEEE (2020)
51. Wang, L., Yan, S., Zhen, J., Liu, Y., Zhang, M., Zhang, G., Zhou, X.: Deep active contours for real-time 6-dof object tracking. In: IEEE/CVF International Conference on Computer Vision (ICCV). pp. 14034–14044 (2023)
52. Wang, P., Liu, L., Liu, Y., Theobalt, C., Komura, T., Wang, W.: Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. In: Ranzato, M., Beygelzimer, A., Dauphin, Y.N., Liang, P., Vaughan, J.W. (eds.) Advances in Neural Information Processing Systems (NeurIPS). pp. 27171–27183 (2021)
53. Wen, B., Mitash, C., Ren, B., Bekris, K.E.: se (3)-tracknet: Data-driven 6d pose tracking by calibrating image residuals in synthetic domains. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 10367–10373. IEEE (2020)
54. Wen, B., Tremblay, J., Blukis, V., Tyree, S., Müller, T., Evans, A., Fox, D., Kautz, J., Birchfield, S.: Bundlesdf: Neural 6-dof tracking and 3d reconstruction of unknown objects. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 606–617 (2023)
55. Wen, B., Yang, W., Kautz, J., Birchfield, S.: Foundationpose: Unified 6d pose estimation and tracking of novel objects. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 17868–17879 (2024)
56. Wu, H., Hu, Z., Li, L., Zhang, Y., Fan, C., Yu, X.: Nefii: Inverse rendering for reflectance decomposition with near-field indirect illumination. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4295–4304 (2023)
57. Yao, Y., Zhang, J., Liu, J., Qu, Y., Fang, T., McKinnon, D., Tsin, Y., Quan, L.: Neilf: Neural incident light field for physically-based material estimation. In: European Conference on Computer Vision (ECCV). pp. 700–716. Springer (2022)

58. Yariv, L., Gu, J., Kasten, Y., Lipman, Y.: Volume rendering of neural implicit surfaces. In: Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., Vaughan, J.W. (eds.) Advances in Neural Information Processing Systems (NeurIPS). vol. 34, pp. 4805–4815. Curran Associates, Inc. (2021), https://proceedings.neurips.cc/paper_files/paper/2021/file/25e2a30f44898b9f3e978b1786dcd85c-Paper.pdf
59. Yu, M., Lu, T., Xu, L., Jiang, L., Xiangli, Y., Dai, B.: Gsdf: 3dgs meets sdf for improved neural rendering and reconstruction. Advances in Neural Information Processing Systems (NeurIPS) **37**, 129507–129530 (2024)
60. Zhang, B., Fang, C., Shrestha, R., Liang, Y., Long, X., Tan, P.: Rade-gs: Rasterizing depth in gaussian splatting. CoRR **abs/2406.01467** (2024)
61. Zhang, J., Yao, Y., Li, S., Liu, J., Fang, T., McKinnon, D., Tsin, Y., Quan, L.: Neilf++: Inter-reflectable light fields for geometry and material estimation. In: IEEE/CVF International Conference on Computer Vision (ICCV). pp. 3601–3610 (2023)
62. 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)
63. Zhao, X., Srinivasan, P.P., Verbin, D., Park, K., Martin-Brualla, R., Henzler, P.: Illuminerf: 3d relighting without inverse rendering. In: Globerson, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J., Zhang, C. (eds.) Advances in Neural Information Processing Systems (NeurIPS). vol. 37, pp. 42593–42617. Curran Associates, Inc. (2024). https://doi.org/10.52202/079017-1349, https://proceedings.neurips.cc/paper_files/paper/2024/file/4b1d9a1fbf7b2a93bea08e18792fe436-Paper-Conference.pdf
64. Zhu, Z.L., Yang, J., Wang, B.: Gaussian splatting with discretized sdf for relightable assets. In: IEEE/CVF International Conference on Computer Vision (ICCV). pp. 25155–25164 (2025)