3DReflecNet: A Large-Scale Dataset for 3D Reconstruction of Reflective, Transparent, and Low-Texture Objects

Supplementary Material

A. Web Interface and Dataset Access

We have developed a web interface, illustrated in Figure 1, to facilitate browsing and retrieval of the 3D instances.

B. Instance Breakdown

To support various downstream tasks, we provide different types of ground truth data to serve as supervised labels. For each synthesis instance, we provide:

- 50 views RGB PNG images of $1,000 \times 1,000$ resolution
- corresponding depth images in exr format
- corresponding normals images in exr format for high precision training
- corresponding mask images
- point cloud file
- Camera internal/external parameters
- Physically-based rendering file (in format of .blend).

For each real-world instance, we provide:

- >60 views RGB PNG images of 1920×1080 (or 3840×2160) resolution
- corresponding mask images
- point cloud file
- Camera internal/external parameters

The design of the instance breakdown can serve different downstream tasks as summarized in Table 1.

C. Near-Field Illumination

While standard HDRI environment maps assume illumination from infinity, they could fail to capture the spatial variations of indoor lighting. To enhance realism, we augment the global environment map with near-field lighting, explicitly simulating local effects by positioning one or two finite-distance point lights on the upper hemisphere. (Figure 2).

D. Assets Generation using 2D image

We generated 3D objects from a dataset of over 5K high-quality real-world captures and synthetic images, using either Hunyuan3D-2.1 [39] or Stable3DGen [47]. The generation process used 50 inference steps and an octree resolution of 320. The resulting objects have an average size of ~ 52 Mb and an average of $\sim 25K$ vertices.

We then performed a manual quality check on the generated assets. While many models, like the electric fan in Figure 3 (left), accurately reproduce the source object's structure, some fail to represent the intended shape faithfully, such as the shoe example in Figure 3 (right). We filtered out these suboptimal results, retaining a final dataset of over 20K high-quality shapes. Figure 15 shows more examples of these selected shapes and materials. Finally, Figure 16 showcases the realistic light reflection behavior on a generated steel asset under various lighting conditions.

E. Image Matching

We provide additional details on the image matching task. Current methods often struggle to deliver satisfactory image matching accuracy for reflective, transparent, or low-texture objects under varying viewpoints, due to their complex and view-dependent appearance characteristics. Figure 4 show cases Eloftr [44] performance under these challenging cases. We also provide quantitative results on our real-world captures. For this, we sampled 100 instances to benchmark the performance of several SOTA image matching methods, with results presented in Table 2.

The overall modest scores shown in Table 2, with the top-performing method only achieving 59.5 at AUC@20°, indicate that robustly matching our real-world instances of challenging materials remains a significant and unsolved problem.

F. Detailed Surface Reconstruction Results

We provide a detailed quantitative comparison of surface reconstruction baselines in Table 3. The results are broken down by our five main material categories to show how performance varies with material complexity.

All methods, including 2DGS, exhibit a clear performance degradation as material complexity increases. The best results are achieved on Diffuse materials, followed by a noticeable drop on Glossy surfaces, and a severe drop on Metallic and Transparent materials. The failure of methods like PGSR and Instant-NGP is particularly evident in the Hausdorff distance, which explodes on non-Lambertian materials, indicating a catastrophic failure to reconstruct large parts of the geometry.

G. Highlight and Specular Reflection Removal

G.1. Preliminary

Highlight removal and specular reflection removal address different types of image artifacts caused by light interactions with surfaces and transparent media. Both problems can be formally described using simplified physical models.

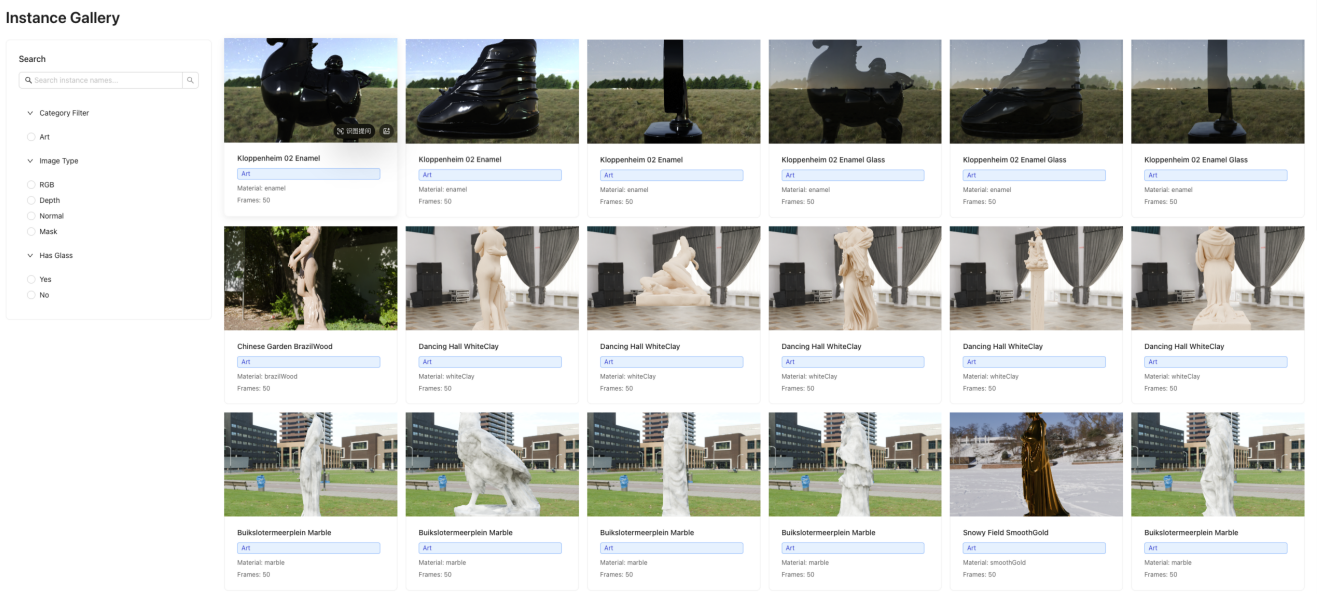

Figure 1. Webpage: User-friendly interface for easy browsing and retrieving 3D instance.

Table 1. Overview of dataset composition and supported downstream applications.

Design	Task
Normals	Normal Estimation [46, 47]; Surface Reconstruction [19, 31]
Depths	Depths Estimation [48]
Masks	Segmentation [33], Detection [5]
(non-)Uniform Lighting	Inverse rendering [24], Highlight Reflection Removal [6, 8]
Multi-view Renderings	Image Matching [7, 36], Structure from Motion [35], 3D Reconstruction [19]
w/ (w/o) Glass Reflection	Specular Reflection Removal [42]

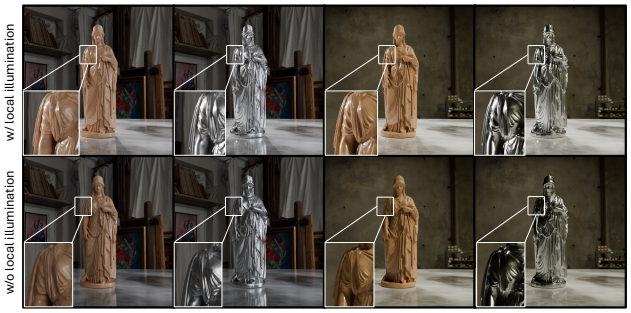

Figure 2. **Synthetic Enhancement for Indoor Scenes.** Top: Local illumination with finite-distance lights. Bottom: Standard infinite-distance HDRI.

Highlight removal focuses on separating the observed image intensity I into its diffuse and specular components:

$$I = I_d + I_s \tag{1}$$

where I_d denotes the diffuse component and I_s the specular component. The diffuse component originates from subsurface scattering and re-emission of light, while the specular component arises from direct reflection of incident light. The magnitude and spatial distribution of I_s depend on surface properties such as roughness and material type.

Table 2. Evaluation of Image Matching on Real-World Capture

Method	**AUC@5°↑**	**AUC@10°↑**	**AUC@20°↑**
SuperPoint+ NN	23.9	33.8	41.2
SuperPoint + SuperGlue	26.2	43.1	51.3
SuperPoint+ LightGlue	25.8	43.5	51.9
LoFTR	27.1	45.9	50.8
AspanFormer	28.3	47.2	51.2
ELoFtr	28.5	47.1	53.7
ROMA	34.3	49.9	59.5

In contrast, specular reflection removal addresses images captured through transparent media, such as glass. In such cases, the observed image I is a mixture of a transmission layer I_t and a reflection layer I_r:

$$I = I_t + I_r \tag{2}$$

Here, I_t corresponds to the scene visible through the trans-

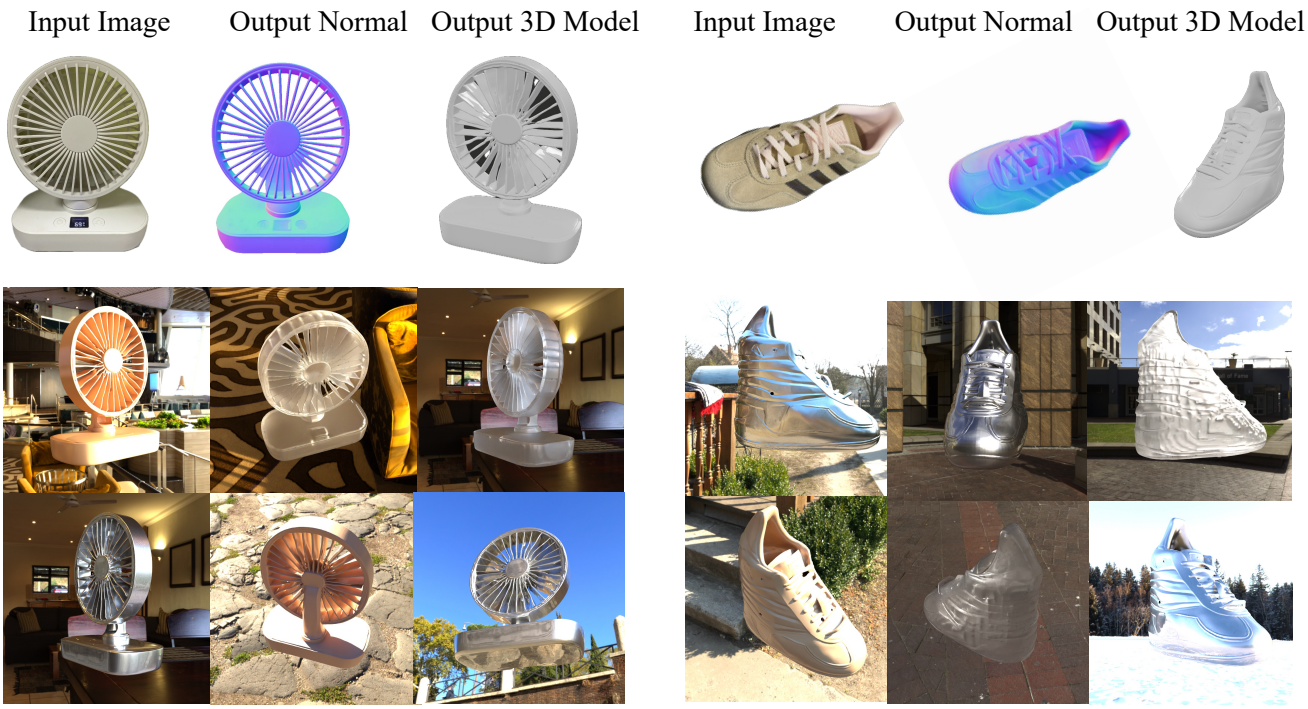

Figure 3. Qualitative Results of 3D Generated Models with Various Materials and Environment Maps. This figure showcases generated models with diverse materials under different lighting. The instance on the left demonstrates a high-quality shape, while the instance on the right shows a failure case that was filtered from our final dataset.

Table 3. Detailed quantitative comparison of surface reconstruction methods on the 3DReflecNet dataset, broken down by material category. 'NGP' refers to Instant-NGP [28]

	Diffuse				**Transparent**				**Metallic**				**Glossy-Textured**				**Glossy-Low-Texture**			
Metric	2DGS	NGP	Neus2	PGSR	2DGS	NGP	Neus2	PGSR	2DGS	NGP	Neus2	PGSR	2DGS	NGP	Neus2	PGSR	2DGS	NGP	Neus2	PGSR
Hausdorff ↓	0.303	1.119	0.872	0.483	2.347	3.214	2.982	6.842	1.862	2.873	2.641	5.923	1.053	1.924	1.632	4.102	1.248	2.105	1.824	4.876
PSNR ↑	35.32	32.96	33.55	34.06	22.76	20.04	21.65	17.95	26.41	23.58	24.22	19.84	31.05	28.10	29.32	24.32	29.87	26.84	27.91	22.10
SSIM ↑	0.960	0.941	0.949	0.954	0.811	0.762	0.789	0.742	0.867	0.821	0.834	0.772	0.928	0.885	0.902	0.821	0.912	0.863	0.884	0.795
LPIPS ↓	0.070	0.098	0.087	0.088	0.190	0.253	0.218	0.347	0.188	0.242	0.212	0.328	0.121	0.179	0.148	0.256	0.138	0.192	0.167	0.284

parent surface, often degraded by refraction and absorption, while I_r results from light reflected off the surface of the medium. These components are often modeled as:

$$I_t = \alpha I_T \tag{3}$$

$$I_r = \beta(I_R * k) \tag{4}$$

where I_T and I_R denote the original transmission and reflection images, respectively; α and β are weighting coefficients; and k is a degradation kernel accounting for blurring or distortion introduced by the reflective surface.

While both tasks involve decomposing a mixed image into multiple layers, they differ in their physical assumptions and application contexts. Highlight removal targets localized reflections on opaque surfaces, whereas specular reflection removal addresses global reflections through transparent materials. Removing specular highlights and reflections enhances image-matching accuracy and, in turn, improves the final 3D reconstruction.

G.2. Evaluation of Reflection Removal on 3DReflecNet

We evaluated four state-of-the-art reflection removal baselines on the 3DReflecNet dataset. For this experiment, we uniformly sampled 1,000 images, applying each method and reporting the average PSNR and SSIM against the ground truth transmission layer. The quantitative results are presented in Table 4.

The results in Table 4 show a clear performance trend, with DSIT achieving the best results (24.07 PSNR / 0.795 SSIM), followed by RRW and DSRNet. These modest scores are comparable to those reported on other challenging real-world datasets. This consistency validates that our experimental setup is effective and that 3DReflecNet serves as a challeng-

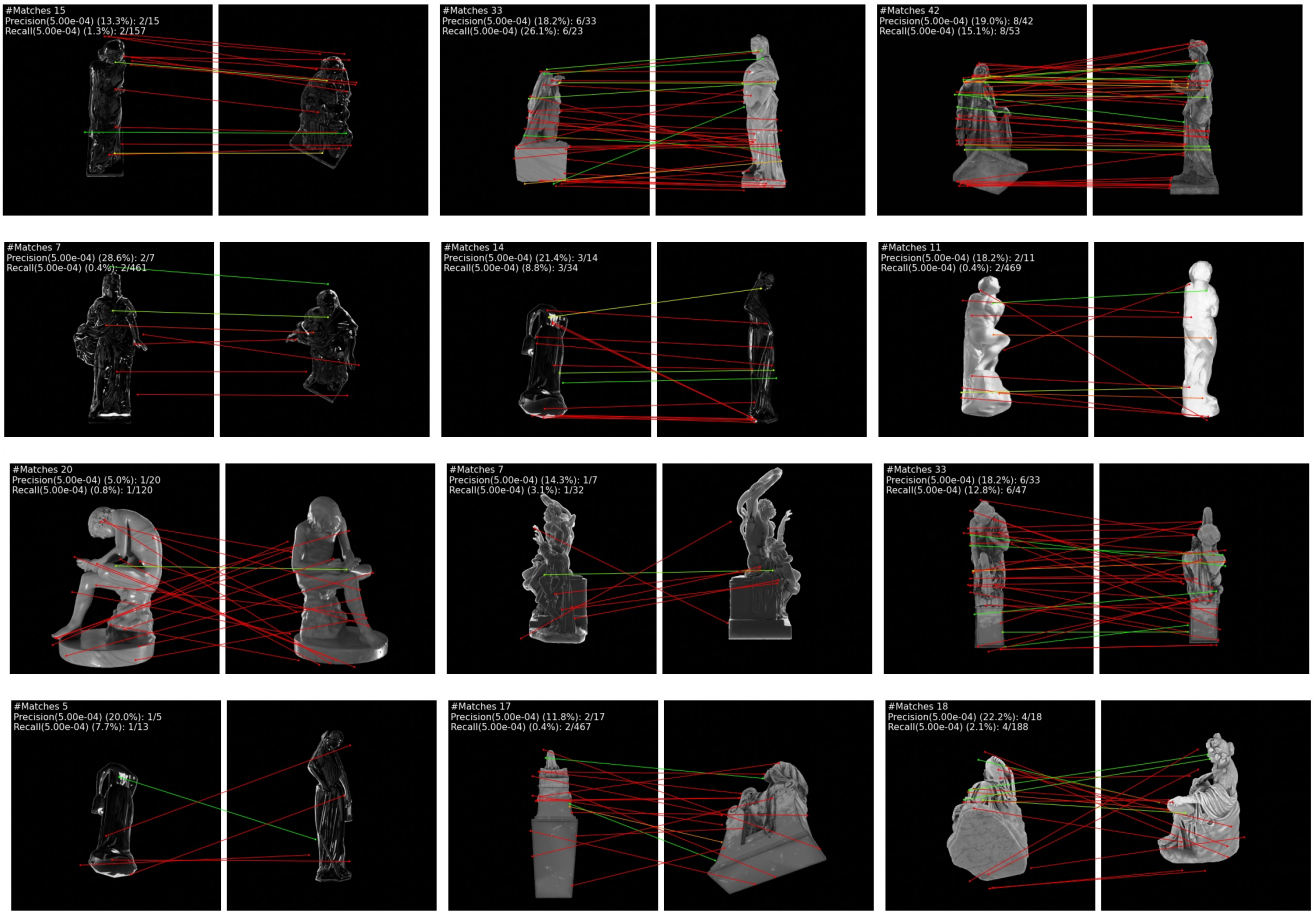

Figure 4. Qualitative results on Efficient Loftr [44] on reflective, transparent Materials.

Table 4. Quantitative comparison of reflection removal baselines on 1,000 images from the 3DReflecNet dataset. Higher is better for both metrics.

Metrics	ERRNet [45]	DSRNet [11]	RRW [52]	DSIT [12]
PSNR (↑)	19.49	22.92	23.11	24.07
SSIM (↑)	0.701	0.782	0.791	0.795

ing and physically-realistic benchmark.

H. Evaluation of NVS and Surface Reconstruction on Real-World Captures

To validate the challenges of our dataset, we benchmarked SOTA methods on our real-world captures. The following tables present the performance for Novel View Synthesis (NVS) and Surface Reconstruction tasks.

NVS Analysis. The NVS results on our real-world data in Table 5 show a clear performance hierarchy, with 2DGS (33.16 PSNR) performing best, followed by Splatfacto (32.07 PSNR). The NeRF-based Instant-NGP (28.12 PSNR) lags significantly behind the Gaussian Splatting methods. These scores, which are lower than those for the synthetic Diffuse category, confirm that our real-world captures, with their complex materials and lighting, pose a significant challenge.

Table 5. Novel View Synthesis performance on the real-world dataset. Metrics: PSNR↑ (LPIPS↓)

Method	**Real-world Dataset**
Instant-NGP [28]	28.12 (0.025)
3DGS [19]	30.99 (0.020)
Splatfacto [37]	32.07 (0.020)
2DGS [13]	33.16 (0.021)

Surface Reconstruction Analysis. The surface reconstruction results in Table 6 highlight the significant challenge our real-world dataset poses for all tested methods. The results

Table 6. Surface Reconstruction performance on the real-world dataset. Metric: Chamfer Distance (↓). Lower is better.

Method	Real-World Dataset
Instant-NGP [28]	0.139
Neus2 [43]	0.132
2DGS [13]	0.105
PGSR [1]	0.207

demonstrates the similar phenomenon to those on the synthetic dataset, indicating that the SOTA methods, designed for simple Lambertian surfaces, are not robust to the complex, non-Lambertian challenges our dataset exposes.

I. Relighting

We evaluated four SOTA relighting methods on the 3DReflecNet dataset. These methods are designed to decompose materials into their intrinsic properties (albedo, roughness, etc.) to allow for rendering under novel lighting conditions. We report the average PSNR and SSIM for novel-light rendering against the ground-truth images.

Table 7. Quantitative comparison of SOTA relighting baselines on the 3DReflecNet dataset.

Metrics	GS-IR [22]	GI-GS [3]	NVDiffrec [29]	TensoIR [16]
PSNR	19.32	20.02	16.98	23.39
SSIM	0.812	0.826	0.795	0.874

The results in Table 7 show that TensoIR achieves the best performance among the baselines, with a PSNR of 23.39. The other Gaussian-based methods, GI-GS and GS-IR, perform moderately, while NVDiffrec struggles significantly. However, the modest scores of all methods, with the top performer failing to exceed 24 dB PSNR, confirm the findings from our main paper: the complex, non-Lambertian materials in 3DReflecNet pose a significant challenge for current SOTA relighting techniques. We thus believe our large-scale dataset of physically-based assets will be a valuable resource for driving future research in this area.

J. Detailed Analysis of Material Parameter Impact

To provide a comprehensive visual reference for the parameter sweep analyzed in the main text, Figure 11 presents the rendered appearance of all 48 unique material configurations tested in our experiments. The 48 combinations are structured into three distinct physical categories. Each category contains 16 variations derived from a 4x4 grid spanning Roughness (0.0, 0.3, 0.6, 0.9) and IOR (1.0, 1.3, 1.6, 1.9):

- Opaque Non-Metals (Dielectrics): (Metallic=0, Transmission=0). 16 combinations.
- Opaque Metals (Conductors): (Metallic=1, Transmission=0). 16 combinations.
- Transparent Non-Metals (Dielectrics): (Metallic=0, Transmission=1). 16 combinations.

The physically implausible combination of (Metallic=1, Transmission=1) was excluded, resulting in the 48 total test cases. This figure allows for a direct visual correlation between a material's appearance and its corresponding reconstruction quality shown in the main paper's analysis.

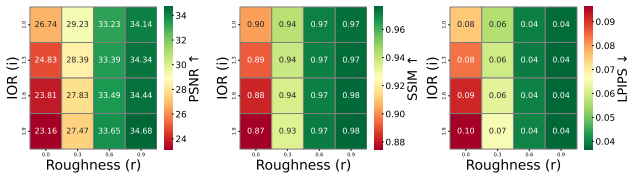

Figure 5. Detailed Impact of Roughness and IOR on Reconstruction Quality for Opaque, Non-Metallic Materials.

The heatmaps in Figure 5 provide a granular analysis of the 16 parameter combinations for opaque, non-metallic materials (Metallic=0, Transmission=0). The three heatmaps plot reconstruction quality metrics against Roughness (x-axis) and Index of Refraction (y-axis). This analysis reveals two key insights:

Roughness is the Dominant Factor for Opaque, Non-Metallic Materials. A strong, consistent trend is visible across all three metrics: reconstruction quality fails catastrophically at low roughness and improves dramatically as roughness increases. At Roughness=0.0, PSNR values are poor, clustering in the 23-27 dB range. As roughness increases to 0.9, the PSNR values improve significantly to the 34-35 dB range. This confirms the hypothesis from the main text: smooth, low-texture surfaces starve the 3DGS algorithm of the high-frequency features needed for multi-view correspondence. As roughness increases, the material's microsurface scatters light more diffusely, which effectively acts as a high-frequency texture that the algorithm can successfully use for matching, drastically reducing failure.

IOR has a Minimal Effect in Opaque, Non-Metallic Cases. In contrast to the strong influence of roughness, the IOR has a very weak, almost negligible, impact on reconstruction quality. For a non-metallic, opaque material, the IOR's primary physical effect is controlling the intensity of specular reflections (the Fresnel effect). Across all metrics, the vertical columns in the heatmaps are nearly uniform in color. For example, at a Roughness of 0.3, the PSNR only varies from 27.47 to 29.23 (a < 2 dB difference) as IOR spans its entire 1.0-1.9 range. At high Roughness=0.9, the IOR has virtually no impact, with PSNR remaining static between 34.14 and 34.68. This is because the high roughness diffuses all reflections, rendering the IOR-driven Fresnel effect imperceptible.

This detailed breakdown confirms that the reconstruction failures in this subset are driven almost entirely by the lack of geometric features on smooth surfaces, not by the view-dependent reflectivity introduced by IOR. This stands in stark contrast to the transparent and metallic cases, where reflectivity and refraction are the primary causes of failure.

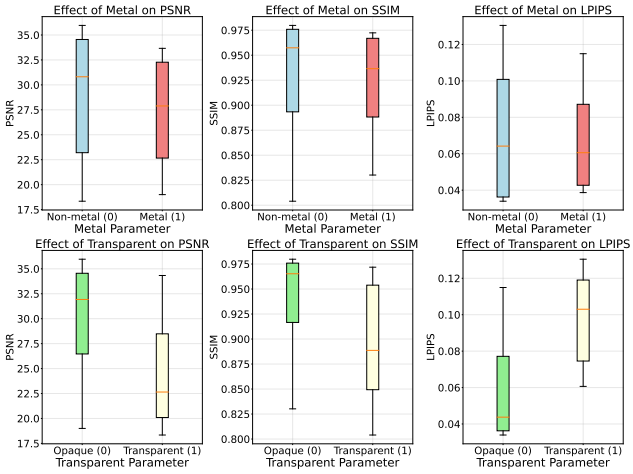

Figure 6. Comparative analysis of reconstruction quality for metallic vs. non-metallic and transparent vs. opaque materials. The box plots show the distribution of PSNR, SSIM, and LPIPS results when aggregating all other material variations.

Analysis of Binary Parameter Effects (Metal & Transparent). While the heatmaps analyze the non-metallic, opaque subset, Figure 6 analyzes the aggregated impact of the binary 'Metal' and 'Transparent' parameters across all 48 configurations.

Metallic materials cause significant pixel-wise failure. Setting 'Metal=1' has a severe negative impact on reconstruction quality, causing a dramatic drop in median PSNR (from 33 dB to 25 dB) and SSIM (from 0.96 to 0.91).Interestingly, this degradation is not reflected in the perceptual LPIPS metric, where the median error for metallic objects (0.062) is slightly better than for non-metallic ones (0.064).

Transparent materials consistently degrade all metrics. Transparency ('Transparent=1') is a more consistent failure case, degrading performance across all metrics. It causes a clear drop in median PSNR (from 30 dB to 28 dB) and SSIM (from 0.95 to 0.92). This negative effect is most pronounced in the perceptual LPIPS metric, where the median error for transparent objects (0.08) is significantly worse than for opaque objects (0.05).

K. A Physically-Based Analysis of Failure Modes in Multi-View 3D Reconstruction

K.1. The Physics of Image Formation: A Ground-Truth Model

To comprehend why certain algorithms fail, one must first establish a ground-truth model of the physical process they attempt to approximate. In computer graphics and physics, the interaction of light with surfaces is meticulously described by the principles of radiometry and physically based rendering. This section establishes a comprehensive and mathematically rigorous model of image formation, which will serve as the physical reality against which the simplified models used in computer vision are compared and critiqued.

Formulate Light in Equilibrium. The cornerstone of physically based rendering is the Rendering Equation, an integral equation independently introduced by Immel et al.,[14] and Kajiya et al.,[18] in 1986. It provides a complete and elegant description of the equilibrium state of light transport in a scene, defining the amount of light leaving any given point on a surface in any given direction. The equation is a statement of energy conservation, asserting that the total light leaving a point is the sum of the light it emits and the light it reflects from all other sources in the environment.

Its canonical form is expressed as:

$$L_o(x,\omega_o) = L_e(x,\omega_o) + \int_\Omega f_r(x,\omega_i,\omega_o) L_i(x,\omega_i)(\mathbf{n}\cdot\omega_i) d\omega_i \tag{5}$$

Each component of this equation has a precise physical meaning:

- $L_o(x,\omega_o)$: The outgoing radiance from a point x on a surface in a specific direction ω_o. Radiance is the radiometric quantity of light energy per unit solid angle per unit projected area, and it is what a camera sensor ultimately measures to form an image.
- $L_e(x,\omega_o)$: The emitted radiance from point x in direction ω_o. This term is non-zero only for surfaces that are light sources themselves. For most objects in a scene, this term is zero.
- $\int_\Omega$: An integral over the unit hemisphere Ω oriented around the surface normal $\mathbf{n}$ at point x. This signifies that to calculate the total reflected light, one must account for all possible incoming light directions from the entire hemisphere above the surface.
- $f_r(x,\omega_i,\omega_o)$: The Bidirectional Reflectance Distribution Function (BRDF). This function is the heart of material appearance, defining the ratio of reflected radiance in the outgoing direction ω_o to the incident irradiance from an incoming direction ω_i.It mathematically describes the intrinsic reflective properties of the material at point x.
- $L_i(x,\omega_i)$: The incident radiance arriving at point x from direction ω_i. This term is what makes the Rendering Equation a global and recursive construct. The light arriving at

point x is simply the outgoing light, L_o, from some other point in the scene that is visible from x along the direction $-\omega_i$. This recursive definition means that the appearance of a single point is dependent on the appearance of every other point in the scene, modeling phenomena like indirect illumination and color bleeding.

- $(\mathbf{n} \cdot \omega_i)$: Lambert's Cosine Law. This is a geometric term representing the dot product between the surface normal $\mathbf{n}$ and the incoming light direction ω_i. It accounts for the fact that a surface receives less light flux per unit area from sources at grazing angles, as the incident energy is spread over a larger area.

The recursive and integral nature of the Rendering Equation reveals a fundamental truth about image formation: it is a global phenomenon. The color of a single pixel is not a purely local property but is the result of a complex interplay of light bouncing throughout the entire scene, converging at that point before traveling to the camera. This global light transport system, which includes inter-reflections between surfaces, is a physical reality that most local, patch-based computer vision algorithms fundamentally fail to model.

K.2. Light Behavior with Complex Materials

A material's electronic response, internal composition, and surface microgeometry dictate the fate of impinging light—whether it is reflected, refracted, transmitted, or absorbed. In rendering and inverse-vision, these processes are commonly expressed via bidirectional scattering functions and Fresnel's laws [17, 26].

K.2.1. Fresnel Reflectance and Refraction

At the interface between media with refractive indices n_1 and n_2, the proportions of reflected and refracted light are governed by Fresnel's equations [17]. For unpolarized light, the reflectance F_r as a function of incident angle θ_i is

$$F_r(\theta_i) = \frac{1}{2}\left[\left(\frac{n_1\cos\theta_i - n_2\cos\theta_t}{n_1\cos\theta_i + n_2\cos\theta_t}\right)^2 + \left(\frac{n_2\cos\theta_i - n_1\cos\theta_t}{n_2\cos\theta_i + n_1\cos\theta_t}\right)^2\right] \tag{6}$$

with θ_t given by Snell's law $n_1 \sin\theta_i = n_2 \sin\theta_t$ [9]. The transmitted fraction T then satisfies energy conservation $F_r + T + A = 1$, where A is absorption.

K.2.2. Microfacet BRDF for Specular Reflection

Metals and glossy dielectrics exhibit specular reflection that varies with surface roughness. The microfacet BRDF [4] is

$$f_r(\omega_i, \omega_o) = \frac{D(h)\, G(\omega_i, \omega_o)\, F_r(\omega_i \cdot h)}{4\cos\theta_i\, \cos\theta_o} \tag{7}$$

where ω_i, ω_o are the incident and exitant directions, h is the half-vector, D the normal distribution function, G the geometric shadowing–masking term, and F_r the Fresnel term. We adopt the Trowbridge–Reitz (GGX) [40] distribution

$$D(h) = \frac{\alpha^2}{\pi\left[(\alpha^2 - 1)\cos^2\theta_h + 1\right]^2} \tag{8}$$

and the Smith–Walter G function [10, 41]. Figure 7 illustrates the process.

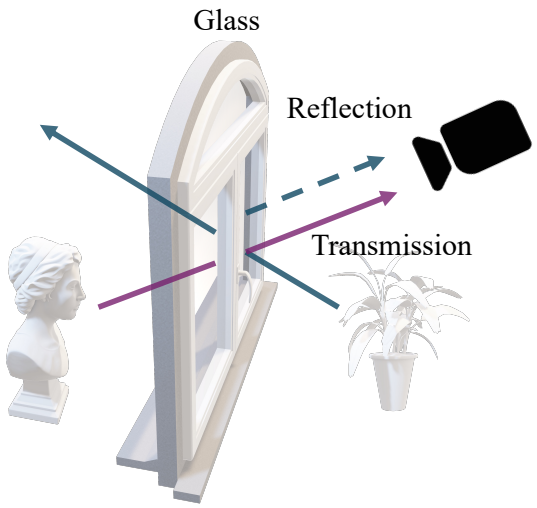

Figure 7. Specular Reflection

K.3. BTDF for Transmission

Transparent dielectrics require a bidirectional transmission distribution function (BTDF) to model refracted light. A common form combines Fresnel-weighted refraction with microfacet masking [41]:

$$f_t(\omega_i, \omega_o) = \frac{(1 - F_r)\, D(h)\, G(\omega_i, \omega_o)\, (n_2/n_1)^2}{4\cos\theta_i\, \cos\theta_o} \tag{9}$$

K.4. Diffuse Scattering and Absorption

Pigmented or rough materials scatter light diffusely. Lambert's law approximates uniform scattering:

$$f_d = \frac{\rho}{\pi} \tag{10}$$

where ρ is albedo. Subsurface scattering in plastics and paints can be modeled via the Kubelka–Munk theory [20] or dipole diffusion [15].

K.5. Foundational Assumptions in Multi-View Reconstruction

Standard reconstruction pipelines, composed of Structure-from-Motion (SfM) and Multi-View Stereo (MVS), are built upon core assumptions that simplify the complex physics of light into a tractable problem.

K.5.1. The Assumption of Photometric Consistency in MVS

The central assumption of MVS is that a true 3D point on a surface will exhibit a similar color across multiple camera

Table 8. Mapping of Material Properties to Violated Algorithmic Assumptions.

Material Type	Primary Phenomenon	Violated Assumption(s)	Consequence
Low-Texture	Lack of high-frequency albedo variation.	Feature Correspondence, Photometric Uniqueness	SfM fails to find matches. MVS cost volume is ambiguous, leading to noisy/flat geometry.
Reflective	View-dependent specular reflection (BRDF).	Photometric Consistency	SfM matching can be unreliable. MVS produces noisy, inaccurate, or incomplete geometry.
Transparent	Refraction of light (Snell's Law), internal reflections.	Photometric Consistency, Linear Light Propagation (Epipolar Geometry)	Complete failure. Photometric matching is impossible. Triangulation is geometrically invalid.

views. MVS algorithms construct a cost volume by comparing image patches between views at hypothesized depths, seeking the depth that minimizes the matching cost (e.g., Sum of Squared Differences). This assumption of view-invariant appearance is, in essence, an implicit assumption of Lambertian reflectance. Any deviation from this ideal diffuse behavior represents a potential violation of this core assumption.

K.5.2. The Assumption of Feature Correspondence in SfM

SfM pipelines operate by detecting and matching sparse, salient feature points (e.g., using SIFT [25]) across multiple views to solve for camera poses. This relies on the assumption that the local appearance of a feature is sufficiently stable across viewpoints to allow for reliable matching. This assumption breaks down under the drastic, non-linear appearance changes caused by strong specular reflections.

K.5.3. The Assumption of Linear Light Propagation in Epipolar Geometry

The entire geometric framework of multi-view reconstruction is predicated on a simple and fundamental assumption: light travels in a straight line from a 3D point to the camera center.Any phenomenon that causes the light path to bend, such as refraction, will invalidate the principles of epipolar geometry and the triangulation methods used to compute 3D structure.

K.6. Conclusion and Future Directions.

The failures of conventional multi-view 3D reconstruction pipelines on reflective, low-texture, and transparent materials are not isolated algorithmic bugs. They are the direct and predictable consequences of a fundamental conflict between the simplified physical models embedded in these algorithms and the complex reality of light transport. The entire SfM-MVS pipeline is built on a set of assumptions that hold only for a small subset of real-world scenes: those that are well-textured, opaque, and largely diffuse. The core argument can be summarized by mapping material properties to the specific assumptions they violate in Table 8.

L. Annotations for Generative 3D Vision Tasks

We extend 3DReflecNet beyond perception tasks by providing detailed textual annotations for each instance, enabling future research in generative 3D vision. This section describes our annotation methodology, format, and provides comprehensive examples to facilitate integration with downstream generation pipelines.

L.1. Annotation Methodology

Our annotation pipeline leverages the Qwen3-VL-30B-A3B-Instruct [38] vision model to generate structured descriptions for each instance in 3DReflecNet. The annotation process captures four key aspects:

1. Detailed Material Descriptions: Comprehensive descriptions of surface properties, including reflectance characteristics, roughness, transparency, and material composition.
2. Lighting Condition Tags: Explicit annotations of lighting setup, including light types, directions, intensities, and environmental illumination.
3. Semantic Instance Descriptions: Object-level descriptions that capture both geometric and appearance properties relevant for 3D generation.

L.2. Annotation Format

Each instance in 3DReflecNet is annotated with structured text following a hierarchical schema. The format includes separate fields for material properties, lighting descriptions, and a natural language generation description which can be used as prompts for downstream tasks.

Figure 8. Input images used for annotation (low/middle/high camera angles).

L.3. Annotation Examples

We provide a concrete example of our annotation pipeline in Figure 8, Figure 9 and Figure 10.

Figure 8 and Figure 9 show the images and complete prompt used to instruct the VLLM, along with a sample *meta.json* file. This input file provides database information, such as *material_name*, and the ground-truth categories.

Figure 10 shows the corresponding *tags.json* file generated by our pipeline. As demonstrated, the process strictly adheres to the prompt's rules: it correctly copies the *category* field from the input and populates the *material_properties*, *environment*, and *description* fields based on the model's visual analysis.

L.4. Integration with Generation Pipelines

Our rich annotations enable seamless integration with various generative 3D vision pipelines. We discuss specific use cases for different generation tasks.

L.4.1. Text-to-3D Generation

The natural language generation prompts can be directly used as conditioning text for diffusion-based 3D generation models [32, 49]. The material and lighting tags enable:

- Material-aware NeRF optimization: Tags guide material parameter initialization and constraints during neural radiance field training
- PBR parameter prediction: Separate channels for roughness, metallic, and normal maps
- Multi-view consistent rendering: Lighting descriptions ensure photometric consistency across views

L.4.2. Text-to-Texture Synthesis

Material property annotations guide texture generation pipelines [2, 34]:

- Roughness and metallic maps: Surface property map synthesis
- Normal map inference: Surface detail generation from descriptions
- Environment-aware baking: Lighting-consistent texture synthesis

L.4.3. Image-to-3D Reconstruction

Lighting descriptions enable advanced reconstruction techniques [23, 47]:

Metric	Value
Total annotated instances	126768
Average description length	268.74 words
Unique material	22
Unique lighting conditions	2700+
Material Complexity Distribution:	
Diffuse	18.2%
Transparent	9.1%
Metallic	22.7%
Glossy-Textured	13.6%
Glossy-Low-Texture	36.4%
Lighting Condition Distribution:	
Indoor - Furnished	8.58%
Indoor - Empty	16.22%
Outdoor - Natural	28.24%
Outdoor - Urban	21.22%
Studio	25.74%

Table 9. Statistics of generation-task annotations in 3DReflecNet. The dataset covers diverse material types and lighting conditions suitable for various generative 3D vision tasks.

- Relighting capability: Material inference for novel lighting conditions
- Lighting-invariant reconstruction: Robust 3D shape recovery under varying illumination

L.4.4. Image Editing and Manipulation

Rich annotations support material-aware image editing [21, 30]:

- Material-consistent inpainting: Preserving material properties during completion
- Lighting-aware object insertion: Matching illumination of inserted objects
- Physical plausibility checking: Validating edits against material-lighting interactions

L.5. Annotation Statistics

Table 9 provides comprehensive statistics of our annotation dataset, demonstrating the scale and diversity of our annotations for generative tasks.

M. Related Works

M.1. Specular Highlight & Reflection Removal

Specular Highlight Removal (SHR) and Single Image Reflection Removal (SIRR) aim to separate interfering light from true scene content. SIRR is a severely ill-posed problem, and modern deep learning approaches are often bottlenecked by an "insufficiency of densely-labeled training data". Recent work like RRW [52] confronts this by creating large-scale,

Asset Image Annotation Prompt

Prompt:

You are a rigorous visual annotator. Please output a concise, structured result **containing only JSON** based on three images of the same instance (low/flat/high angles).

- The model is placed on a marble surface.
- The camera height ratio for the three angles are 0.3 / 1.0 / 1.5 (relative to the model height).
- category.main/sub must be **strictly copied** from meta.json; no reclassification is allowed.
- Only judge "material appearance attributes" and "environment"; do not output database fields (hasGlass, transparent, isGenerated, material_name).
- All values must be selected from the provided enums or value ranges.

Note: Output Fields (Only These!)

- instance_id: string (from meta.json or parameters)
- category: { main: string, sub: string } (strictly copied from meta.json; reclassification is prohibited)
- material_properties:
  - glossiness: "matte"—"semi-gloss"—"glossy"—"mirror"
  - roughness: number (0.0-1.0)
  - reflectivity: "none"—"weak"—"medium"—"strong"
  - texture_scale: "fine"—"medium"—"coarse"
  - anisotropy: "none"—"weak"—"strong"
  - metallic_hint: "non-metal"—"mixed"—"metallic"
- environment:
  - indoor_outdoor: "indoor"—"outdoor"—"unknown"
  - light_type: "natural"—"artificial"—"mixed"—"unknown"
  - light_intensity: "weak"—"medium"—"strong"—"unknown"
  - key_light_direction: "left"—"right"—"front"—"back"—"top"—"ambient"—"unknown"
  - shadow_hardness: "soft"—"medium"—"hard"—"unknown"
- description: string (briefly describe the appearance of the instance in English)

Note: Category Fields to Be Strictly Copied!

- category.main = "{cat_main}"
- category.sub = "{cat_sub}"

The following is meta.json (for reference only; do not use its database fields for judgment):


```
{
    "instance_id": "7db354ab-34a0-4634-b4e3-cf100a4736c9",
    "main_category": "culture",
    "sub_category": "sculpture",
    "model_name": "Roman_Statue.blend_Zenobia in Chains statue",
    "material_name": "cleanSnow",
    "env_name": "quarry_cloudy_4k.exr",
    "hasGlass": false,
    "isGenerated": false,
    "transparent": false
}
```


Figure 9. The prompt for generating description for an asset

aligned real-world datasets. Architecturally, the field has evolved from location-aware models [6] to dual-stream networks, such as the interactive transformers in DSIT [12]. The related field of SHR also relies on deep learning, highlighting the importance of "leveraging large-scale synthetic data" for generalization [8].

M.2. Relighting

Modern 3D reconstruction with Neural Radiance Fields [27] excels at view synthesis but entangles geometry, materials,

Generated Annotation Output (tags.json)

This is the resulting *tags.json* file generated by the annotation pipeline, corresponding to the input from Figure 9.


```
{
  "instance_id": "7db354ab-34a0-4634-b4e3-cf100a4736c9",
  "category": {
    "main": "culture",
    "sub": "sculpture"
  },
  "material_properties": {
    "glossiness": "matte",
    "roughness": 0.3,
    "reflectivity": "weak",
    "texture_scale": "fine",
    "anisotropy": "none",
    "metallic_hint": "non-metal"
  },
  "environment": {
    "indoor_outdoor": "outdoor",
    "light_type": "natural",
    "light_intensity": "medium",
    "key_light_direction": "top",
    "shadow_hardness": "soft"
  },
  "description": "A classical marble sculpture of a robed female figure,
  standing on a white marble platform. The statue has a matte, light gray
  finish with fine surface details. It is positioned outdoors in a quarry-like
  environment with a body of water and rocky hills in the background.
  The lighting is soft and diffused, suggesting an overcast sky."
}
```


Figure 10. The structured *tags.json* output from our annotation pipeline. This includes the strictly copied category, the VLLM-inferred material and environment properties, and the natural language description.

and lighting, which hinders relighting. This ”baked-in” problem spurred research into disentangling these properties. Early works like NeRFactor [51], and PhySG [50] factorized the implicit field but remained computationally expensive and often limited to low-frequency lighting. To overcome this, two explicit strategies emerged. First, Munkberg et al., [29] jointly optimized an explicit triangular mesh, materials, and all-frequency lighting using a differentiable rasterizer and DMTet. Second, the paradigm shifted to 3D Gaussian Splatting, GS-IR [22] adapted inverse rendering to this efficient representation to decompose physical properties. The most recent works, such as GI-GS [3], now address the limitations of initial 3DGS methods by explicitly modeling global illumination, often using screen-space path tracing to separate direct and indirect lighting.

N. More Qualitative Examples

We provide additional qualitative examples in this section. Figure 13 showcases various shapes, while Figure 14 details objects with different materials. For our 3D-generated instances, Figure 15 shows diverse shapes and materials, and Figure 16 displays a steel asset under various lighting conditions. Finally, Figure 12 presents more real-world capture instances.

References

[1] Danpeng Chen, Hai Li, Weicai Ye, Yifan Wang, Weijian Xie, Shangjin Zhai, Nan Wang, Haomin Liu, Hujun Bao, and Guofeng Zhang. Pgsr: Planar-based gaussian splatting for efficient and high-fidelity surface reconstruction. *IEEE Transactions on Visualization and Computer Graphics*, 31(9): 6100–6111, 2025. 5

[2] Dave Zhenyu Chen, Yawar Siddiqui, Hsin-Ying Lee, Sergey

Tulyakov, and Matthias Nießner. Text2tex: Text-driven texture synthesis via diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 18558–18568, 2023. 9

[3] Hongze Chen, Zehong Lin, and Jun Zhang. Gi-gs: Global illumination decomposition on gaussian splatting for inverse rendering. *arXiv preprint arXiv:2410.02619*, 2024. 5, 11

[4] Robert L. Cook and Kenneth E. Torrance. A reflectance model for computer graphics. *ACM Transactions on Graphics*, 1(1): 7–24, 1982. 7

[5] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superpoint: Self-supervised interest point detection and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 224–236, 2018. 2

[6] Zheng Dong, Ke Xu, Yin Yang, Hujun Bao, Weiwei Xu, and Rynson W.H. Lau. Location-aware single image reflection removal. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5017–5026, 2021. 2, 10

[7] Johan Edstedt, Qiyu Sun, Georg Bökman, Mårten Wadenbäck, and Michael Felsberg. Roma: Robust dense feature matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19790–19800, 2024. 2

[8] Gang Fu, Qing Zhang, Lei Zhu, Chunxia Xiao, and Ping Li. Towards high-quality specular highlight removal by leveraging large-scale synthetic data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12857–12865, 2023. 2, 10

[9] John E. Greivenkamp. *Field Guide to Geometrical Optics*. SPIE Press, 2004. 7

[10] Eric Heitz. Understanding the masking–shadowing function in microfacet-based brdfs. *Journal of Computer Graphics Techniques*, 3(2):24–78, 2014. 7

[11] Qiming Hu and Xiaojie Guo. Single image reflection separation via component synergy. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 13138–13147, 2023. 4

[12] Qiming Hu, Hainuo Wang, and Xiaojie Guo. Single image reflection separation via dual-stream interactive transformers. *Advances in Neural Information Processing Systems*, 37: 55228–55248, 2024. 4, 10

[13] Binbin Huang, Zehao Yu, Anpei Chen, Andreas Geiger, and Shenghua Gao. 2d gaussian splatting for geometrically accurate radiance fields. In *ACM SIGGRAPH 2024 conference papers*, pages 1–11, 2024. 4, 5

[14] David S. Immel, Michael F. Cohen, and Donald P. Greenberg. A radiosity method for non-diffuse environments. In *Proceedings of the 13th Annual Conference on Computer Graphics and Interactive Techniques*, page 133–142, New York, NY, USA, 1986. Association for Computing Machinery. 6

[15] Henrik W. Jensen, Steve R. Marschner, Marc Levoy, and Pat Hanrahan. A practical model for subsurface light transport. In *Proceedings of SIGGRAPH*, pages 511–518, 2001. 7

[16] Haian Jin, Isabella Liu, Peijia Xu, Xiaoshuai Zhang, Songfang Han, Sai Bi, Xiaowei Zhou, Zexiang Xu, and Hao Su. Tensoir: Tensorial inverse rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 165–174, 2023. 5

[17] Deane B Judd. Fresnel reflection of diffusely incident light. 1942. 7

[18] James T Kajiya. The rendering equation. In *Proceedings of the 13th annual conference on Computer graphics and interactive techniques*, pages 143–150, 1986. 6

[19] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023. 2, 4

[20] Paul Kubelka and Franz Munk. Ein beitrag zur optik der farbanstriche. *Zeitschrift für Technische Physik*, 12:593–601, 1931. 7

[21] Shanglin Li, Bohan Zeng, Yutang Feng, Sicheng Gao, Xiuhui Liu, Jiaming Liu, Lin Li, Xu Tang, Yao Hu, Jianzhuang Liu, et al. Zone: Zero-shot instruction-guided local editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6254–6263, 2024. 9

[22] Zhihao Liang, Qi Zhang, Ying Feng, Ying Shan, and Kui Jia. Gs-ir: 3d gaussian splatting for inverse rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21644–21653, 2024. 5, 11

[23] Minghua Liu, Ruoxi Shi, Linghao Chen, Zhuoyang Zhang, Chao Xu, Xinyue Wei, Hansheng Chen, Chong Zeng, Jiayuan Gu, and Hao Su. One-2-3-45++: Fast single image to 3d objects with consistent multi-view generation and 3d diffusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10072–10083, 2024. 9

[24] Yuan Liu, Peng Wang, Cheng Lin, Xiaoxiao Long, Jiepeng Wang, Lingjie Liu, Taku Komura, and Wenping Wang. Nero: Neural geometry and brdf reconstruction of reflective objects from multiview images. *ACM Transactions on Graphics (ToG)*, 42(4):1–22, 2023. 2

[25] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004. 8

[26] Alexander I Lvovsky. Fresnel equations. *Encyclopedia of Optical Engineering*, 27:1–6, 2013. 7

[27] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 10

[28] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM transactions on graphics (TOG)*, 41(4):1–15, 2022. 3, 4, 5

[29] Jacob Munkberg, Jon Hasselgren, Tianchang Shen, Jun Gao, Wenzheng Chen, Alex Evans, Thomas Müller, and Sanja Fidler. Extracting triangular 3d models, materials, and lighting from images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8280–8290, 2022. 5, 11

[30] Hyelin Nam, Gihyun Kwon, Geon Yeong Park, and Jong Chul Ye. Contrastive denoising score for text-guided latent diffusion image editing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9192–9201, 2024. 9

[31] Michael Oechsle, Songyou Peng, and Andreas Geiger. Unisurf: Unifying neural implicit surfaces and radiance fields for multi-view reconstruction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5589–5599, 2021. 2

[32] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022. 9

[33] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 2

[34] Elad Richardson, Gal Metzer, Yuval Alaluf, Raja Giryes, and Daniel Cohen-Or. Texture: Text-guided texturing of 3d shapes. In *ACM SIGGRAPH 2023 conference proceedings*, pages 1–11, 2023. 9

[35] Noah Snavely, Steven M Seitz, and Richard Szeliski. Photo tourism: exploring photo collections in 3d. In *ACM siggraph 2006 papers*, pages 835–846. 2006. 2

[36] Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. Loftr: Detector-free local feature matching with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8922–8931, 2021. 2

[37] Matthew Tancik, Ethan Weber, Evonne Ng, Ruilong Li, Brent Yi, Terrance Wang, Alexander Kristoffersen, Jake Austin, Kamyar Salahi, Abhik Ahuja, et al. Nerfstudio: A modular framework for neural radiance field development. In *ACM SIGGRAPH 2023 conference proceedings*, pages 1–12, 2023. 4

[38] Qwen Team. Qwen3 technical report, 2025. 8

[39] Tencent Hunyuan3D Team. Hunyuan3d 2.1: From images to high-fidelity 3d assets with production-ready pbr material, 2025. 1

[40] TS Trowbridge and Karl P Reitz. Average irregularity representation of a rough surface for ray reflection. *Journal of the optical society of America*, 65(5):531–536, 1975. 7

[41] Bruce Walter, Steve Marschner, Hongsong Li, and Kenneth Torrance. Microfacet models for refraction through rough surfaces. In *Eurographics Symposium on Rendering*, pages 195–206, 2007. 7

[42] Tianfu Wang, Mingyang Xie, Haoming Cai, Sachin Shah, and Christopher A Metzler. Flash-split: 2d reflection removal with flash cues and latent diffusion separation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5688–5698, 2025. 2

[43] Yiming Wang, Qin Han, Marc Habermann, Kostas Daniilidis, Christian Theobalt, and Lingjie Liu. Neus2: Fast learning of neural implicit surfaces for multi-view reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3295–3306, 2023. 5

[44] Yifan Wang, Xingyi He, Sida Peng, Dongli Tan, and Xiaowei Zhou. Efficient loftr: Semi-dense local feature matching with sparse-like speed. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 21666–21675, 2024. 1, 4

[45] Kaixuan Wei, Jiaolong Yang, Ying Fu, David Wipf, and Hua Huang. Single image reflection removal exploiting misaligned training data and network enhancements. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8178–8187, 2019. 4

[46] Chongjie Ye, Lingteng Qiu, Xiaodong Gu, Qi Zuo, Yushuang Wu, Zilong Dong, Liefeng Bo, Yuliang Xiu, and Xiaoguang Han. Stablenormal: Reducing diffusion variance for stable and sharp normal. *ACM Transactions on Graphics (TOG)*, 43 (6):1–18, 2024. 2

[47] Chongjie Ye, Yushuang Wu, Ziteng Lu, Jiahao Chang, Xiaoyang Guo, Jiaqing Zhou, Hao Zhao, and Xiaoguang Han. Hi3dgen: High-fidelity 3d geometry generation from images via normal bridging. *arXiv preprint arXiv:2503.22236*, 3, 2025. 1, 2, 9

[48] Xinyi Ye, Weiyue Zhao, Tianqi Liu, Zihao Huang, Zhiguo Cao, and Xin Li. Constraining depth map geometry for multi-view stereo: A dual-depth approach with saddle-shaped depth cells. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17661–17670, 2023. 2

[49] Xin Yu, Yuan-Chen Guo, Yangguang Li, Ding Liang, Song-Hai Zhang, and Xiaojuan Qi. Text-to-3d with classifier score distillation. *arXiv preprint arXiv:2310.19415*, 2023. 9

[50] Kai Zhang, Fujun Luan, Qianqian Wang, Kavita Bala, and Noah Snavely. Physg: Inverse rendering with spherical gaussians for physics-based material editing and relighting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5453–5462, 2021. 11

[51] 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. 11

[52] Yurui Zhu, Xueyang Fu, Peng-Tao Jiang, Hao Zhang, Qibin Sun, Jinwei Chen, Zheng-Jun Zha, and Bo Li. Revisiting single image reflection removal in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 25468–25478, 2024. 4, 9

Figure 11. Materials of various parameters show different physical phenomenon. The parameters is in format of $< m, r, i, t >$

Figure 12. More real-world capture instances, including semi-transparent, reflective, and low-texture objects.

Figure 13. The synthetic objects of various shapes

Figure 14. The object with the same shape but made of different materials under identical lighting condition

Figure 15. Various shapes of generated 3D assets made of different materials

Figure 16. Generated 3D assets made of Steel material under various lighting condition