

REBUTTAL

Anonymous authors

Paper under double-blind review

We thank the reviewers for their constructive feedback and positive recognition of ReLi3D’s contributions. Reviewers acknowledge that *‘to my knowledge, the suggested approach is the first which jointly reconstructs mesh, PBR, and environment (HDR), all in a feedforward manner and at impressive speeds’* **FiCS**, constituting *‘a significant contribution’* **FiCS**. ReLi3D is recognized as *‘the first unified end-to-end system that reconstructs complete 3D geometry, spatially-varying materials, and environment illumination from sparse multi-view images in under one second’* **BeZw**, with *‘fast relightable 3D generation’* **7ZmD** that *‘significantly enhances the practicality and usability of 3D generation systems’* **7ZmD**.

The architectural design is appreciated: *‘the two-path feed-forward framework jointly reconstructs geometry, materials, and illumination under multi-view constraints, showing a clear and coherent design’* **7UYe**, and *‘the idea of fusing arbitrary number of views with one “hero” view and other views with latent mixing (what the authors call “cross-view feature fusion”) is novel and insightful’* **FiCS**. The training approach using *‘Monte Carlo integration with MIS for training supervision’* **7UYe** leads to *‘more consistent and realistic reconstructions’* **7UYe**, while *‘self-supervised learning on real data’* **7ZmD** represents *‘an important advancement over prior works’* **7ZmD**.

Results are described as *‘convincing’* **FiCS**, with *‘strong quantitative results’* **7ZmD** showing *‘large improvements over baseline methods across multiple benchmarks’* **7ZmD**, and the *‘writing is clear and easy to follow’* **BeZw**. We address the reviewers’ concerns and questions in the following sections.

REAL-WORLD PERFORMANCE AND DATASET GENERALIZATION

Real-world Evaluation - **7ZmD 7UYe**

We demonstrated real-world performance beyond the teaser in Appendix Fig. 7, which includes challenging UCO3D Liu et al. (2024) captures with motion blur and cluttered backgrounds. These examples show the benefit of our multi-view setting (e.g., recovering the front of the bear given a second view) and improved material prediction as lighting aligns with ground truth. We will move representative samples to the main paper and reference them in Sec. 5.3 if wished by the reviewers. We further evaluate on real-world data from Stanford ORB Kuang et al. (2024) (Tbl. I) and include even harder real-world example in Fig. I.

Method	Stanford ORB								
	3D				Image			Basecolor	
	CD↓	FS@0.1↑	FS@0.2↑	FS@0.5↑	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑
SF3D	0.152	0.512	0.769	0.954	17.75	0.891	0.111	18.52	0.865
SPAR3D	0.165	0.488	0.751	0.940	17.10	0.886	0.113	17.80	0.857
Trellis	0.152	0.561	0.782	0.948	17.13	0.888	0.112	—	—
Hunyuan3D	0.141	0.571	0.801	0.960	16.96	0.877	0.110	21.37	0.872
ReLi3D (Ours)	0.116	0.608	0.856	0.980	18.68	0.907	0.098	24.21	0.891
Hunyuan3D (2 Views)	0.134	0.588	0.809	0.967	16.91	0.876	0.108	21.42	0.872
Hunyuan3D (4 Views)	0.136	0.579	0.810	0.966	16.83	0.877	0.108	21.46	0.873
ReLi3D (Ours) (2 Views)	0.104	0.654	0.888	0.986	19.74	0.913	0.089	25.01	0.896
ReLi3D (Ours) (4 Views)	0.094	0.718	0.906	0.989	20.84	0.919	0.082	25.33	0.900
ReLi3D (Ours) (8 Views)	0.089	0.745	0.914	0.991	21.21	0.921	0.080	25.50	0.901
ReLi3D (Ours) (16 Views)	0.089	0.749	0.914	0.990	21.29	0.921	0.080	25.58	0.902

Table I: Real-world Evaluation on Stanford ORB. Quantitative evaluation on Stanford ORB dataset showing 3D reconstruction, image quality, and basecolor material prediction performance. Our method outperforms baselines across all metrics and improves with more input views.

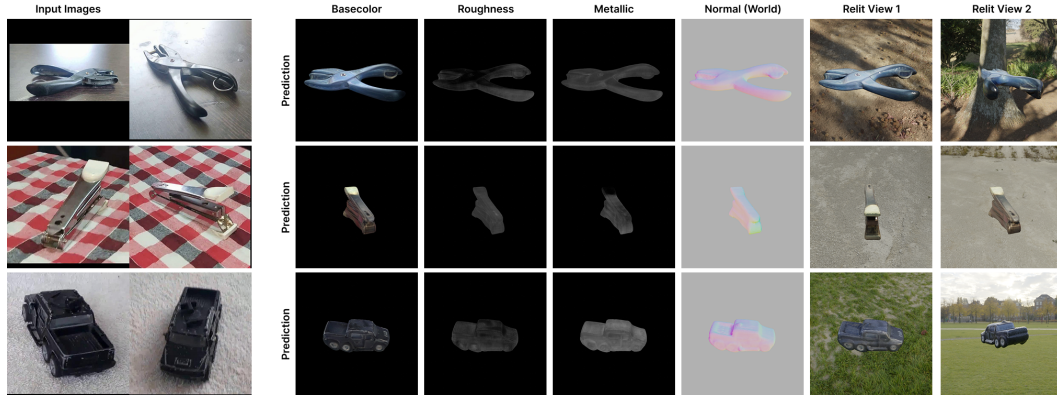


Figure I: Real-world material prediction. Material maps (albedo, roughness, metallic, normal) for real-world objects from UCO3D dataset on very challenging settings, strong reflections and blur. Our method is still able to make a rough prediction and faithfully separates metallic and non-metallic materials.

UCO3D Liu et al. (2024) is RGB-only, and naively training on such data can lead to entangled appearance cues. Our mixed-domain training addresses this: synthetic PBR data provide direct supervision for materials and RENI++ Gardner et al. (2023) latent codes (Sec. 5.1), anchoring the decomposition in physically correct signals. Real UCO3D data are used only with image space reconstruction losses through our Monte Carlo renderer. We additionally apply a demodulation regularizer that biases the environment toward neutral white lighting when no RENI++ ground truth is available. We filter UCO3D using the dataset’s reconstruction and camera-quality scores, discarding sequences with poor cameras, masks, or severe motion blur. See section C.2 and Fig. I for details.

Texture Detail and Perceived Blur - 7ZmD

Our goal is 0.3s reconstruction on a single H100, constraining model size to a max triplane resolution of $(3 \times 40 \times 384 \times 384)$. As noted in Sec. 5.5, current blur results primarily from this resolution and the DINOv2 Oquab et al. (2023) fine-tuning bottleneck, not the disentanglement framework. Higher-capacity variants (e.g., frozen VGGT features and higher triplane resolution (Fig. IV)) yield sharper textures but compromise other aspects. While diffusion models may produce higher frequency texture, they often hallucinate unfaithful details (e.g., Fig. V) and require significantly longer inference times (~ 120 s).

Complex Objects & Materials - 7ZmD 7UYe

We extended evaluation to: (i) complex, multi-material objects from the Blender Shiny dataset (Fig. II), already quantified in the main paper, and (ii) material-focused metrics on Stanford ORB Kuang et al. (2024) (Tbl. I). These results confirm our spatially varying PBR prediction generalizes to complex geometries and real materials. Regarding transparent objects (7UYe): while our density-based NeRF Mildenhall et al. (2021) pre-training handles transparency, explicit mesh reconstruction of transparent surfaces remains an open research challenge and outside our current scope.

MULTI-VIEW DESIGN, HERO VIEW, AND FUSION

Hero View Selection & Sensitivity - 7UYe FiCS 7ZmD

The hero view is the query stream for the cross-conditioning transformer. In Tbls. 1 and 2, the hero view is selected uniformly at random, ensuring our reported metrics reflect robust performance independent of viewpoint choice, unlike methods relying on canonical frontal views. To test sensitivity, we compared random selection against fixed frontal-view selection (Tbl. II). Results show only marginal differences, with slight perceptual gains for random views likely due to parallax information from side views. We will clarify this protocol in Sec. 4.1.1.

Multi-view Improvements & Saturation - 7ZmD

Performance saturation in Tbl. 2 stems from coverage saturation and random view sampling. Once

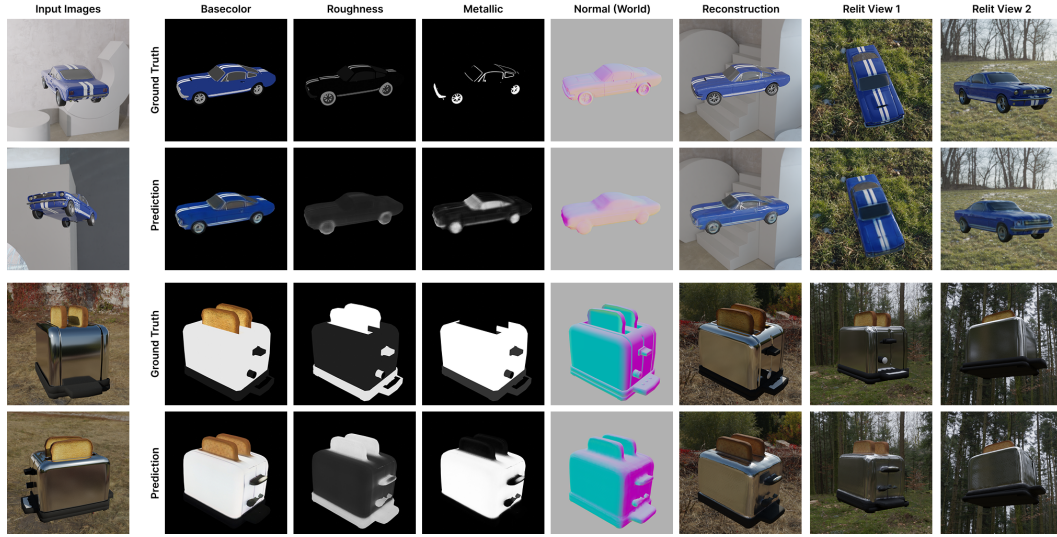


Figure II: Varying materials and complex objects. Results on the Blender Shiny dataset showing spatially varying PBR material prediction on complex multi-material objects. The figure shows predicted basecolor, roughness, metallic, and normal maps, along with relit renderings in novel environments.

Method	Polyhaven Subset						
	3D				Image		
	CD↓	FS@0.1↑	FS@0.2↑	FS@0.5↑	PSNR↑	SSIM↑	LPIPS↓
ReLi3D (Random Hero)	0.102	0.697	0.883	0.982	20.25	0.919	0.083
ReLi3D (Frontal Hero)	0.123	0.641	0.840	0.965	19.06	0.909	0.095

Table II: Hero view selection sensitivity. Comparison of metrics using random hero view selection versus always selecting the most frontal view on the Polyhaven dataset.

surface coverage is sufficient (4-8 views), additional random views often provide redundant information rather than new constraints, leading to marginal gains and minor metric oscillations.

Cross-View Fusion Architecture - 7UYe

Our cross-view fusion enables multi-view inputs within memory and runtime constraints. We need a single shared triplane conditioned on all views while running in under a second on a single GPU.

There are three generic fusion strategies: (1) Input fusion, where images are merged before the transformer; (2) Mid level fusion, as in our cross-conditioning transformer with a hero view; (3) Output fusion, where each view predicts its own triplane and a separate network fuses them afterwards.

Input fusion risks discarding information before the model can reason about occlusions and consistency. Output fusion requires a separate backbone per view plus an additional fusion module, which is significantly more memory intensive and essentially a different architecture. Our cross-conditioning design with a hero view and cross-attention (i) keeps view-specific information until late, (ii) produces a single shared triplane, and (iii) remains trainable within our single-GPU memory budget.

Ablating cross-view fusion in a strict sense would require building and training a completely different family of architectures (e.g., early input fusion or full attention mid fusion or late triplane fusion) at ReLi3D scale, which is not a toggle-style ablation but a separate method. Instead, we provide an implicit ablation already in the paper: for N=1 input view, the model reduces to a pure hero-view transformer with no cross-view context. The improvements when moving from 1 to multiple views in Tbls. 1 and 2 are a direct consequence of adding cross-view fusion and demonstrate its benefit for geometry and relighting quality.

ARCHITECTURE CHOICE AND ABLATIONS VIA PRIOR WORK

Two Path vs Single Path Architecture - **BeZw FiCS**

Our architecture consists of a single unified network with two branches: one branch predicts geometry+materials and the other branch predicts illumination. Both branches share the same cross-view fused triplane features and are evaluated in one forward pass, optimized jointly through the same Monte Carlo rendering objective. This is what we refer to as our “two-path” design. Reviewer **FiCS** contrasts “multiple paths vs a single path,” and Reviewer **BeZw** explicitly suggests an ablation where geometry+materials and environment illumination are predicted by two independent networks instead of this unified model. Conceptually, we agree that this is an interesting question, but we believe that, at the scale of ReLi3D, it is not a simple or particularly informative ablation, for two reasons:

1. Fairness and resources (BeZw): Two independent networks would almost double parameters (as the triplane transformer needs to be cloned), memory, and runtime compared to our unified model. At our current resolution and batch size, this would exceed GPU memory constraints. Reducing each network’s capacity to keep the total parameter count fixed would artificially handicap both networks and still require significantly more training time. In both cases, it is unclear what a fair comparison should look like.

2. Coupling vs. post-hoc fusion (BeZw): In our unified design, illumination and geometry+materials are disentangled inside one coupled system. Both branches see the same fused features and are constrained by the same rendering loss. Splitting this into two independent networks requires an additional communication mechanism (e.g., passing predicted PBR maps to a separate illumination network, or vice versa). Designing and training such a two-stage pipeline is essentially a new method, requiring choices about how to represent and pass information between networks and how to keep gradients stable through both stages. Similar problems are well known in works that predict geometry and texture with separate networks and then struggle to fuse them again, we avoid this complexity by design.

SPAR3D as Ablation - **FiCS**

The method SPAR3D Huang et al. (2025), which we compare against and significantly outperform, effectively corresponds to ablating our internal two-branch illumination design. In SPAR3D, the environment is predicted directly from a triplane head without a dedicated illumination branch. We show in Fig. 4, Fig. III and Tbl. IV that our unified two-branch architecture enables much more accurate illumination prediction than SPAR3D, despite using the same RENI++ Gardner et al. (2023) latent space. This supports the necessity of our illumination branch in practice. Also see our new quantitative results Tbl. IV.

SF3D as Ablation - **BeZw**

In the original SF3D Boss et al. (2024), which we build upon, materials are predicted externally as global parameters, and the authors already report that they need carefully chosen priors (e.g., β -distributions) to avoid collapse and still do not obtain correct material prediction (see Fig. 5). While this is not an illumination experiment, it supports the point that parameters which must be disentangled from a single image (geometry, materials, lighting) are better handled in a unified architecture like ours, where they are optimized jointly under a physically based rendering objective.

A full “two independent networks” variant would either (a) break the single-GPU regime central to ReLi3D, or (b) require substantial architectural redesign beyond a straightforward ablation. As detailed in Appendix B, our current model trains in three stages of 60,000 steps on a single H100 GPU, corresponding to several GPU days for one full run. Training an additional fullscale two-network variant during the rebuttal period would roughly double this budget and is beyond our available compute and time resources.

Training Stage Ablations - **BeZw**

We ablated all training stages. Our progressive pipeline transitions from volumetric rendering through spherical Gaussian approximation stages (128 $\rightarrow$ 256 $\rightarrow$ 512 Gaussians) to full Monte Carlo integration, as detailed in Appendix B.2. Tbl. III reports the share of the total improvement (Phase 1 $\rightarrow$ Full MC) contributed by each intermediate stage. The Gaussian stages with larger batch sizes explain 70–80% of the 3D coverage gains (CD/FS), confirming they mainly stabilize geometry before expensive rendering. The Monte Carlo stage accounts for the majority of remaining im-

improvements in material disentanglement (basecolor, roughness, metallic). The 512-Gaussian stage provides the sweet spot for geometry+runtime, while the final MC finetuning sharpens material maps and relighting without regressing 3D accuracy.

Due to the high computational cost of training ReLi3D, we focused our ablation budget on the most critical component: the MC+MIS renderer. As shown in Tbl. 3, removing MC rendering and using only a simplified renderer (“– MC-Render”) drops image PSNR from 19.92 to 17.54 dB and slightly degrades geometry metrics, confirming that physically based rendering is central to our disentanglement. Training full alternative architectures (e.g., completely separate nets for geometry and illumination or a different fusion backbone) at this scale would require several additional GPU-weeks, which is beyond our current resources (see section Two Path vs Single Path Architecture). Instead, we provide lighter-weight ablations (single-view vs multi-view) that directly target the claimed contributions. We also view SF3D Boss et al. (2024) and SPAR3D Huang et al. (2025) (e.g., illumination comparison) as direct ablations of our design.

Method	3D Coverage Share (%)	Image Quality Share (%)	Basecolor Share (%)	Roughness Share (%)	Metallic Share (%)
Phase 2 (256) vs Phase 1 (128)	20.2	6.8	22.6	6.3	7.7
Phase 3 (512) vs Phase 2 (256)	70.3	50.0	23.9	31.3	41.0
MC (Full) vs Phase 3 (512)	9.5	43.2	53.5	62.4	51.3

Table III: Training stage contribution analysis. Average share of the total improvement from Phase 1 (128 Gaussians) to the full Monte Carlo stage that is attributable to each intermediate stage. Columns aggregate the metrics shown in Table 1: (1) 3D coverage (CD and FS@0.05–0.5), (2) image quality (PSNR, SSIM, LPIPS), (3) basecolor (PSNR, SSIM, LSSIMSE), (4) roughness (PSNR, SSIM, RMSE), and (5) metallic (PSNR, SSIM, RMSE). Early Gaussian stages mainly expand 3D coverage, while the Monte Carlo refinement sharpens PBR material disentanglement.

ILLUMINATION MODELING AND DISENTANGLEMENT

Illumination Prior & Alternative Representations - FiCS

Our framework is compatible with alternative lighting representations: we use spherical Gaussian approximations in the intermediate training stage (Appendix B.2) before switching to Monte Carlo rendering with RENI++ Gardner et al. (2023) envmaps. In those stages, we train with a low frequency Gaussian representation and observe that it fails to capture sharp highlights and directional suns, leading to worse relighting metrics as shown in Tbl. 3.

RENI++ Gardner et al. (2023) provides a compact but high-frequency representation critical for photorealistic relighting and accurate material and lighting separation. While nothing in our architecture prevents using SH or Gaussians, we found RENI++ to be the best trade-off between expressiveness and efficiency. We choose this compact representation to fit our memory limitations, expanding into a more memory intensive representation (e.g., ENV Map HDR prediction) would not be possible with our constraints. See the SPAR3D Huang et al. (2025) ablation discussion, which highlights that not only the representation choice matters but also how it is interfaced.

Background Masking Strategy - FiCS

This occlusion strategy is part of the training strategy for the illumination transformer. As detailed in Sec. 4.1.3, we encode mask image pairs (M_i, I_i) with a dedicated DINOv2-small Oquab et al. (2023) encoder and then apply stochastic background masking. During training, we randomly occlude background pixels in a subset of views. This forces the network to solve two complementary tasks: when background pixels are visible, it can read lighting directly from the environment, when they are masked, it must infer lighting from indirect cues in object reflections and shading.

This is equivalent to training with a mixture of fully visible and fully masked backgrounds, implemented as an augmentation. Empirically, this strategy yields more robust lighting estimates in real-world scenes where backgrounds are often partially cropped, saturated, or noisy.

Quantitative Evaluation of Illumination Disentanglement - 7UYe

Fig. 4 compares our predicted HDR environment maps to ground truth and to SPAR3D Huang et al. (2025), which also predicts RENI++ Gardner et al. (2023) latents, showing that ReLi3D recovers sharper, higher-contrast envmaps with correctly localized light sources. Tbl. IV provides quantitative results comparing ReLi3D, SPAR3D, and a DiffusionLight Phongthawee et al. (2023) baseline on the Polyhaven+HDRI dataset. While DiffusionLight Phongthawee et al. (2023) and ReLi3D are on

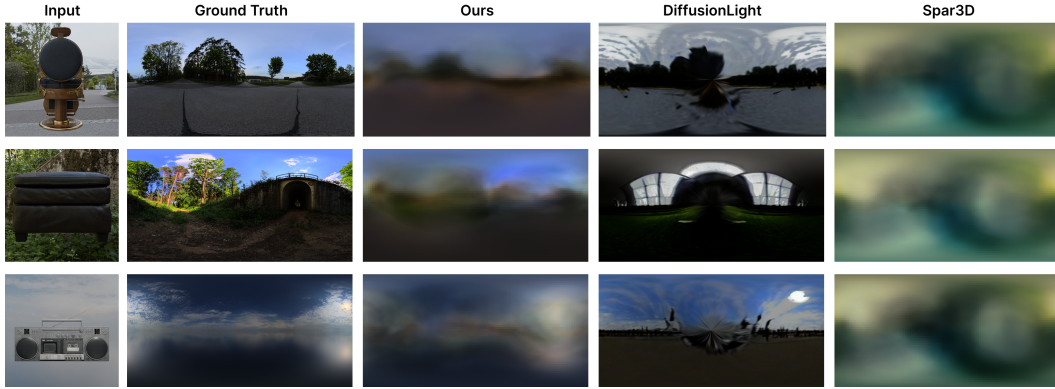


Figure III: Illumination Comparison. Comparison of illumination prediction results between DiffusionLight, SPAR3D, and our method (ReLi3D). Predicted environments vary vastly while ours mimics the ground truth shape and color, DiffusionLight hallucinates a completely different environment, SPAR3D fails.

par for the image PSNR metric, ReLi3D is vastly faster and allow for multi-view input. Qualitatively Fig. III shows that ReLi3D achieves predicting correct horizon lines and colors while DiffusionLight completely hallucinates different environments, even indoor although the input is a simple outdoor scene. SPAR3D fails in every setting.

Time (s)↓			PSNR↑		
Diffusion Light	ReLi3D	SPAR3D	Diffusion Light	ReLi3D	SPAR3D
21.46	0.34	0.33	20.93	20.88	17.10

Table IV: Quantitative evaluation of illumination disentanglement. Comparison of environment map prediction and relighting quality on Polyhaven+HDRI dataset.

GENERATED INPUTS, ROBUSTNESS, AND FAILURE MODES

Composability with Image Generative Models - FiCS

ReLi3D assumes known camera poses and physically plausible materials, which are often violated by generated images. Diffusion models frequently contain non-physical highlights or impossible reflections. We tested ReLi3D on (i) single images synthesized by text-to-image models with estimated camera poses and (ii) built an unposed version with frozen VGGT features and higher triplane resolution (Fig. IV). Single-image inputs generally work well when the estimated pose is accurate (e.g from DUST3R Wang et al. (2024)), but serverly bad pose estimation leads to blur artifacts. While our pose free version has better texture quality, it has other limitations, we leave this for future work. We also tested on multi-view sequences from an off-the-shelf multi-view diffusion model. Multi-view generations sometimes degrade performance due to pose and appearance inconsistencies. Pairing generated multi-view images with proxy 3D reconstructions from a multi-view diffusion pipeline would allow adapting ReLi3D specifically to this regime, a direction worth exploring in future work.

Limitations and Failure Modes - FiCS ZUYe

ReLi3D has failure modes. As briefly discussed in Sec. 5.5, the main failure mode arises when illumination falls outside the RENI++ Gardner et al. (2023) latent distribution or, especially in scenes with multiple extremely bright, localized light sources, but this is a limitation of the RENI++ prior and not of our method. We explored failure cases for our method in Fig. V and find that strong selfshadowing can lead to baked in lighting (see shark fin) in the material maps and dark scenes make basecolor prediction very hard (see Rhino). However, we want to point out that in both cases we still outperform Hunyuan3D Zhao et al. (2025), which is a strong baseline.



Figure IV: Generative input evaluation. ReLi3D performance on single images from diffusion models, shows good results with slight blur due to incorrect pose estimation. We further show results on a pose free version with frozen VGGT features and higher triplane resolution, having better texture quality but other limitations. (Object relit in different illumination then input images)

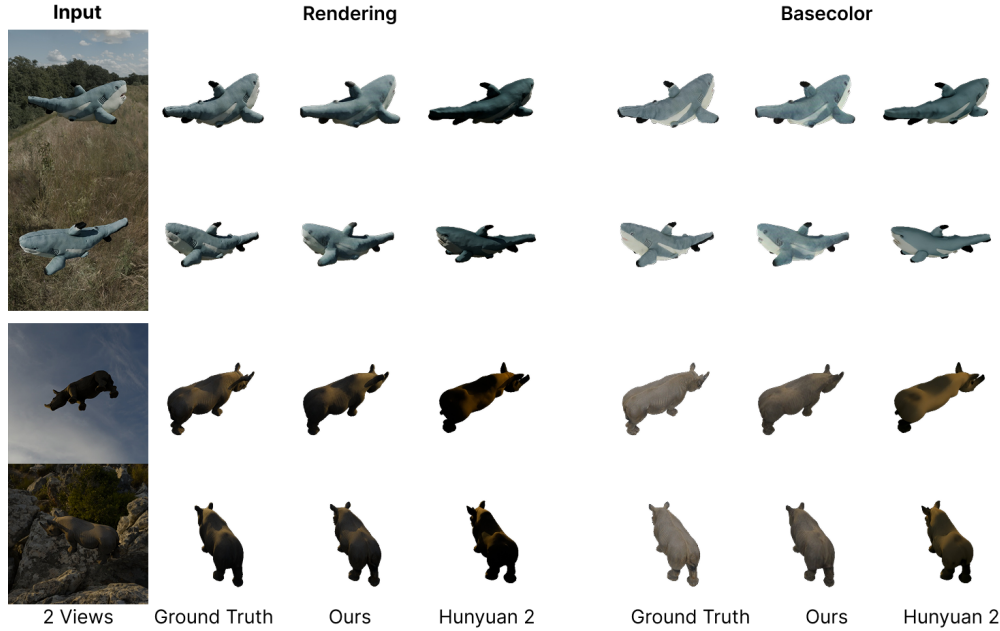


Figure V: Failure cases. Failure cases showing challenges with baked in lighting for objects with strong self shadowing (fin of shark) and basecolor prediction difficulties in dark scenes (Rhino). Comparison includes results from Hunyuan3D and our method.

EVALUATION PROTOCOL AND REMAINING NOTES

Fair Comparison Protocol and Mesh Alignment - BeZw

Baselines often produce meshes in arbitrary canonical spaces. To ensure fair comparison, we apply rigid ICP alignment to the ground truth before evaluating image metrics (Tbels. 1 & 2). ReLi3D's predictions are naturally aligned, so this step primarily aids baselines, highlighting a useful feature of our method for robotic use cases.

Additional Updates

We updated Fig. 6 with higher resolution images (FiCS) and added comprehensive training details and loss definitions to Appendix B.2 (BeZw).

REFERENCES

- Mark Boss, Zixuan Huang, Aaryaman Vasishta, and Varun Jampani. Sf3d: Stable fast 3d mesh reconstruction with uv-unwrapping and illumination disentanglement. *arXiv preprint*, 2024.
- James AD Gardner, Bernhard Egger, and William AP Smith. Reni++ a rotation-equivariant, scale-invariant, natural illumination prior. *arXiv preprint arXiv:2311.09361*, 2023.
- Zixuan Huang, Mark Boss, Aaryaman Vasishta, James M Rehg, and Varun Jampani. Spar3d: Stable point-aware reconstruction of 3d objects from single images. 2025.
- Zhengfei Kuang, Yunzhi Zhang, Hong-Xing Yu, Samir Agarwala, Elliott Wu, Jiajun Wu, et al. Stanford-orb: a real-world 3d object inverse rendering benchmark. *Advances in Neural Information Processing Systems*, 36, 2024.
- Xingchen Liu, Piyush Tayal, Jianyuan Wang, Jesus Zarzar, Tom Monnier, Konstantinos Tertikas, Jiali Duan, Antoine Toisoul, Jason Y. Zhang, Natalia Neverova, Andrea Vedaldi, Roman Shapovalov, and David Novotny. Uncommon objects in 3d. In *arXiv*, 2024.
- 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*, 2021.
- Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.
- Pakkapon Phongthawee, Worameth Chinchuthakun, Nontaphat Sinsunthithet, Aditya Raj, Varun Jampani, Pramook Khungurn, and Supasorn Suwajanakorn. Diffusionlight: Light probes for free by painting a chrome ball. *arXiv preprint arXiv:2303.13009*, 2023.
- Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jérôme Revaud. Dust3r: Geometric 3d vision made easy. *arXiv preprint arXiv:2312.14132*, 2024.
- Zibo Zhao, Zeqiang Lai, Qingxiang Lin, Yunfei Zhao, Haolin Liu, Shuhui Yang, Yifei Feng, Mingxin Yang, Sheng Zhang, Xianghui Yang, et al. Hunyuan3d 2.0: Scaling diffusion models for high resolution textured 3d assets generation. *arXiv preprint arXiv:2501.12202*, 2025.