

Digital Twin Catalog: A Large-Scale Photorealistic 3D Object Digital Twin Dataset

Zhao Dong¹ Ka Chen^{1,*} Zhaoyang Lv^{1,*} Hong-Xing Yu^{2,*} Yunzhi Zhang^{2,*} Cheng Zhang^{1,*}
 Yufeng Zhu^{1,*} Stephen Tian² Zhengqin Li¹ Geordie Moffatt¹ Sean Christofferson¹ James Fort¹
 Xiaqing Pan¹ Mingfei Yan¹ Jiajun Wu² Carl Yuheng Ren¹ Richard Newcombe¹

¹Meta Reality Labs Research ²Stanford University

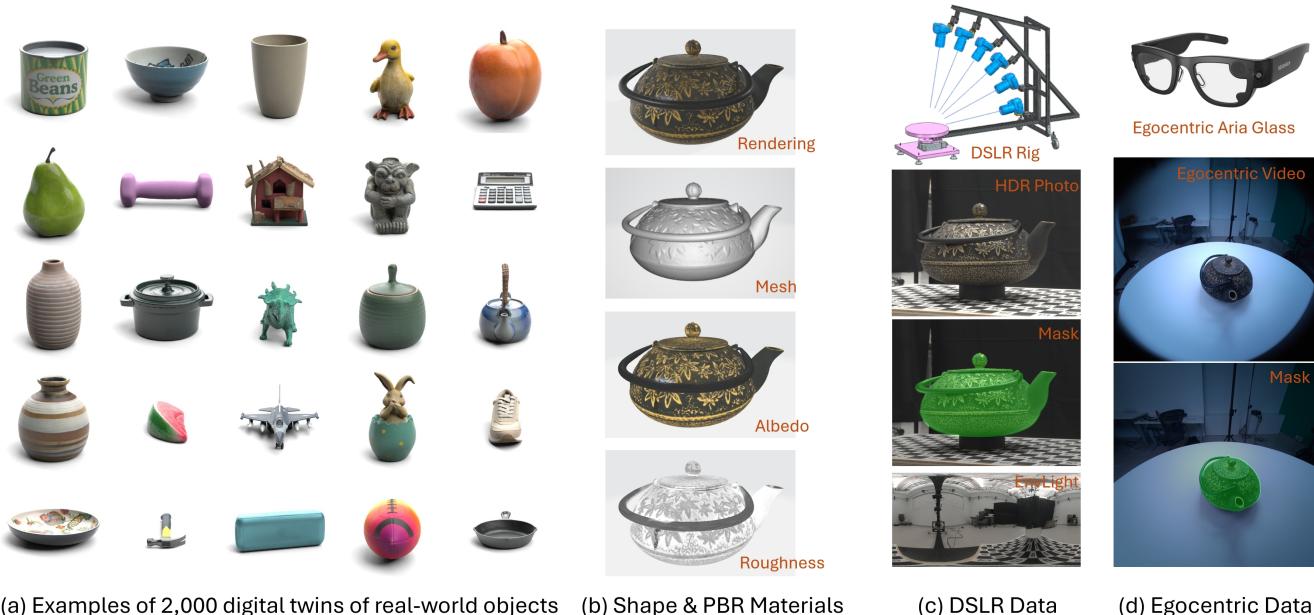


Figure 1. The Digital Twin Catalog (DTC) dataset comprises 2,000 digital twins of physical-world objects (a), characterized by millimeter-level geometric accuracy and photorealistic PBR materials (b). DTC includes evaluation data captured using both DSLR cameras and egocentric Aria glasses, featuring captured images with precise foreground object masks and environment lighting for relighting evaluation.

Abstract

We introduce Digital Twin Catalog (DTC), a new large-scale photorealistic 3D object digital twin dataset. A digital twin of a 3D object is a highly detailed, virtually indistinguishable representation of a physical object, accurately capturing its shape, appearance, physical properties, and other attributes. Recent advances in neural-based 3D reconstruction and inverse rendering have significantly improved the quality of 3D object reconstruction. Despite these advancements, there remains a lack of a large-scale, digital twin quality real-world dataset and benchmark that can quantitatively assess and compare the performance of different reconstruction methods, as well as improve reconstruction quality through

training or fine-tuning. Moreover, to democratize 3D digital twin creation, it is essential to integrate creation techniques with next-generation egocentric computing platforms, such as AR glasses. Currently, there is no dataset available to evaluate 3D object reconstruction using egocentric captured images. To address these gaps, the DTC dataset features 2,000 scanned digital twin-quality 3D objects, along with image sequences captured under different lighting conditions using DSLR cameras and egocentric AR glasses. This dataset establishes the first comprehensive real-world evaluation benchmark for 3D digital twin creation tasks, offering a robust foundation for comparing and improving existing reconstruction methods. The DTC dataset is already released at <https://www.projectaria.com/datasets/dtc/> and we will also make the baseline evaluations open-source.

* Authors contributed equally and are listed in alphabetical order.

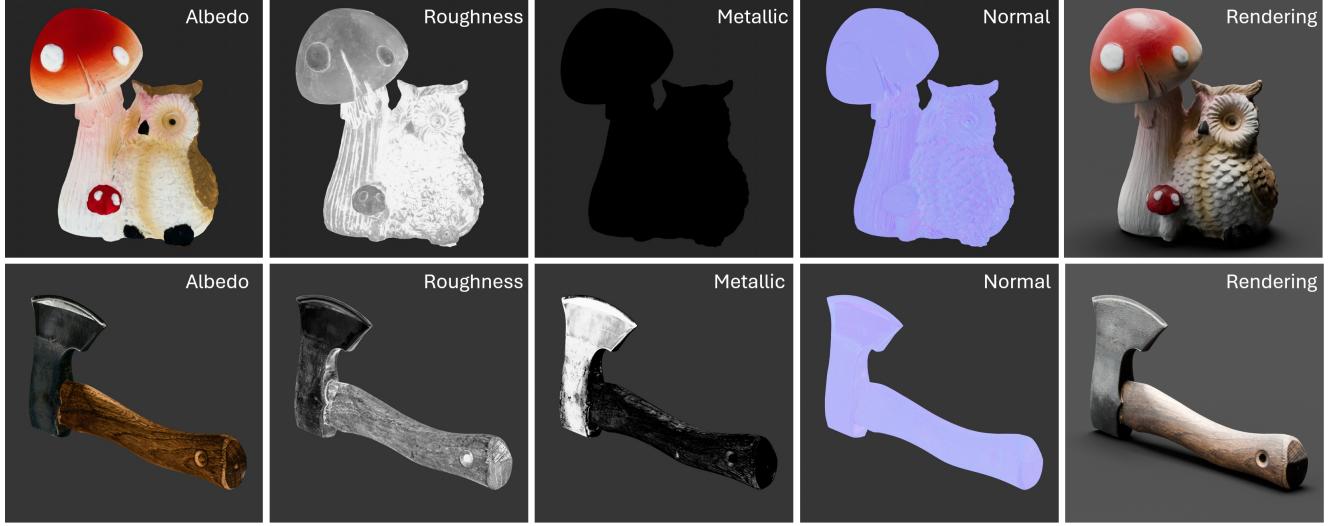


Figure 2. Example DTC models with photorealistic PBR materials.

1. Introduction

A digital twin of a 3D object is a highly detailed, virtually indistinguishable representation of a physical object, capturing its shape, appearance, physical properties and other attributes with precision. Such a digital twin enables visualization, analysis, and interaction as if it were the real object, supporting simulation, automation, and real-world problem-solving across a wide range of applications in AR/VR [3, 28], spatial/contextual AI [2], and robotics [23, 43]. As fundamental properties of an object, its shape and appearance form the basis for recognizing and interpreting the 3D object, enabling identification, manipulation, and realistic rendering. Recovering these attributes has long been a foundational topic in computer vision and graphics, inspiring extensive research in 3D reconstruction and inverse rendering. Recent breakthroughs in neural-based representation and reconstruction techniques, such as NeRF [52] and 3D Gaussian splatting (3DGS) [37], have significantly elevated the quality of novel view synthesis (NVS) to photorealistic levels. Many subsequent works [35, 62] integrate neural reconstruction with physically-based inverse rendering, enabling relightable appearances. Furthermore, leveraging priors from large reconstruction models (LRMs) [30], high-quality shape and appearance reconstruction can now be achieved with as few as one to four views [41, 69, 82].

Despite the rapid advancements in 3D object reconstruction, one question remains: *does the reconstruction quality truly meet the standard of a digital twin, where virtual representations are indistinguishable from reality?* This digital twin standard demands both **highly accurate shape matching** and **photorealistic appearance across different lighting**, which present significant acquisition challenges for real-world objects. Existing object-centric datasets for 3D recon-

struction or inverse rendering have focused on either dataset size [18] or quality of specific aspects [16, 21, 34, 46, 61, 73], often sacrificing comprehensive fidelity and limiting their application scope. This trade-off has led to a lack of datasets that fully satisfy the digital twin criteria, hindering current 3D reconstruction methods from achieving digital twin fidelity. To bridge this gap, we developed the Digital Twin Catalog (DTC) dataset, comprising 2,000 scanned 3D object models (Fig. 1(a)), each with millimeter geometry accuracy and photorealistic PBR materials (Fig. 1(b), Fig. 2).

In addition to 3D digital twin models, the DTC dataset includes evaluation data designed to support 3D object reconstruction research. This evaluation data features multi-view image sequences with precise foreground object masks and environment lighting information for relighting evaluation. Traditionally, high-quality HDR images captured with modern DSLR cameras have been the standard for 3D reconstruction research. Looking ahead, we encourage the integration of 3D reconstruction research with next-generation human-centric computing platforms, such as ego-centric AR glasses, aiming to democratize 3D reconstruction techniques and empower everyone to effortlessly create 3D digital twins. To this end, alongside DSLR-captured evaluation data (Fig. 1(c)), the DTC dataset also provides ego-centric evaluation data captured using Project Aria glasses (<https://www.projectaria.com>) (Fig. 1(d)).

The DTC dataset offers extensive opportunities for advancing research in object digital twin creation. We provide a benchmark for state-of-the-art 3D object reconstruction and inverse rendering methods. These benchmarks evaluate performance across novel view synthesis (NVS), shape reconstruction, and relightable appearance reconstruction. We further provide the evaluation of novel view synthesis meth-

Table 1. Comparison with existing object-centric inverse rendering datasets. *Objaverse [18] consists of both synthetic objects and real scans, only part of which contain PBR materials.

Dataset	# Objects	Real	Scene Type	Multi-view	Shape	PBR Mat.	Relit Image	Lighting	Egocentric Cap.
ShapeNet-Intrinsics [61]	31K	X	synthetic	✓	✓	X	✓	✓	X
NeRD Synthetic [11]	3	X	synthetic	✓	✓	✓	✓	✓	X
ABO [16]	8K	X	synthetic	✓	✓	✓	✓	✓	X
MIT Intrinsics [25]	20	✓	studio	✓	X	X	X	X	X
DTU-MVS [34]	80	✓	studio	✓	✓	X	X	X	X
Objaverse [18]	818K	(✓)*	studio	✓	✓	(✓)*	X	X	X
DiLiGenT-MV [42]	5	✓	studio	✓	✓	X	X	✓	X
ReNe [65]	20	✓	studio	✓	X	X	X	✓	X
OpenIllumination [46]	64	✓	studio	✓	X	✓	✓	✓	X
GSO [21]	1030	✓	studio	✓	✓	X	X	X	X
Lombardi <i>et al.</i> [47]	6	✓	in-the-wild	X	✓	X	✓	✓	X
NeRD Real [11]	4	✓	in-the-wild	✓	X	✓	✓	X	X
NeROIC [38]	3	✓	in-the-wild	✓	X	✓	✓	X	X
Oxholm <i>et al.</i> [56]	4	✓	in-the-wild	✓	✓	X	✓	✓	X
OmniObject3D [73]	6k	✓	in-the-wild	✓	✓	X	X	X	X
Stanford Orb [39]	14	✓	in-the-wild	✓	✓	✓	✓	✓	X
DTC (ours)	2k	✓	in-the-wild	✓	✓	✓	✓	✓	✓

ods using the egocentric aligned DTC data. Additionally, we explore the dataset’s potential in downstream robotics applications by assessing its effectiveness in training robotic policies for pushing and grasping tasks in simulation. These benchmarks and applications provide valuable insights, highlight existing challenges, and uncover promising directions for future research in 3D digital twin creation.

2. Related Work

We provide a comparison of our DTC dataset to existing object-centric datasets in Table 1. We provide the largest 3D dataset with PBR materials and real world multi-view recordings with digital twin counterparts. We further provide digital twin aligned egocentric recordings, the first of their kind in the egocentric domain. We will discuss the related datasets and methods they can empower as follows.

3D Digital Twin Datasets Existing 3D digital twin datasets with PBR materials often serve as ground truth for evaluating inverse rendering results. Early efforts [6, 25] provide small-scale intrinsic image of real objects and do not provide shape or PBR material information. Synthetic datasets [11, 16, 45, 61, 70, 72] are widely used for evaluation but do not represent the complexity in a real world environment. For datasets that contain real objects, [18, 34, 42, 47, 56, 65], the reconstruction quality can vary, which leads to imprecise evaluations. Table 1 provides a comparison to the previous work in this domain. Compared to Objaverse [18, 19], which is a collection of existing 3D models with only a small subset containing PBR materials with varying quality, we offer a high quality collection of 3D object data that is also aligned with real world recordings. Compared to OmniObject3D [73], the DTC models

provide higher-quality shape and additional PBR materials that are necessary for high quality inverse rendering. We offer the largest quantity of 3D object models compared to all counterparts in various tasks. For real-world evaluation, Stanford-ORB [39] was the prior largest inverse rendering benchmark with in-the-wild lighting. In contrast, we provide more object diversity and higher quality for each object model. The Aria Digital Twin dataset [58] was the first dataset to provide digital twin aligned environments for the scenes and recorded using egocentric device. However, their scene environments are limited and the contained object ground truths inside do not have high quality geometries with PBR materials.

Object Reconstruction & Inverse Rendering. Using object-centric multi-view images as input, early object reconstruction methods focused on estimating individual object properties, such as shape from shading [5, 9, 83], material acquisition [44, 45, 56, 57, 74], and lighting estimation [67, 78]. Some approaches also aimed to recover reflectance and illumination assuming known object shapes [47, 48]. Inverse rendering, which seeks to invert the rendering equation [36], estimates an image’s intrinsic components—geometry, materials, and lighting. The advent of differentiable renderers [14, 15, 49] enabled full-fledged inverse rendering methods to simultaneously recover all these properties for object reconstruction [50].

Neural volumetric representations such as Neural Radiance Fields (NeRFs) [52] and the like [7, 11–13, 59, 79, 84] encode geometry and appearance as volumetric densities and radiance with a Multi-Layer Perceptron (MLP) network, and render images using the volume rendering equation [51]. 3D-GS [37] introduces 3D Gaussian primitives and rasterization and its following-up variants [31] demonstrates high quality

geometry prediction as well.

Other surface-based representations [29, 53, 62, 66, 71, 80, 81, 85] extract surfaces as the zero level set, for instance, of a signed distance function (SDF) or an occupancy field [55], allowing them to efficiently model the appearance on the surface with an explicit material model, such as bidirectional reflectance distribution functions (BRDFs). This also enables modeling more complex global illumination effects, such as self-shadows. Most of these methods focus on per-scene optimization and require dense multiple views as input. Recently, researchers have incorporated learning-based models, distilling priors from large training datasets for fast inference on limited test views [8, 10, 33, 45, 61, 70, 72, 86].

In this work, we provide the DTC dataset with real-world object recordings that can serve as the benchmark to evaluate object-centric inverse rendering tasks. We evaluate representative baselines from existing work.

3. Digital Twin Catalog: A Large-Scale Photorealistic 3D Object Digital Twin Dataset

3.1. Dataset Composition

Our DTC dataset contains: (1) **2,000** scanned 3D object models, featuring millimeter geometric accuracy relative to their physical counterparts, along with a full set of photorealistic PBR material maps (Fig. 1), (2) **100** DSLR-captured evaluation data of **50** objects under different lighting conditions, and (3) **200** egocentric Aria-captured evaluation data of **100** objects with both **active** and **casual** observation modes.

3.2. Creation of 3D Object Models

Utilizing the state-of-the-art industrial 3D object scanner [1], we selected 2,000 physical-world objects spanning 40 LVIS [27] categories, carefully chosen to ensure both category diversity and compatibility with the scanner's capabilities. As illustrated in Fig. 3(a)-(c), the scanner features a fixed lighting-camera setup within an upper-hemisphere dome, equipped with 8 structured lights for geometry scanning, and 29 spotlights and 29 cameras for material acquisition. During the scanning process, the object is placed on a central holder, and its pose can be adjusted with multi-round scanning to achieve a complete 360-degree scan. For our dataset, each object typically undergoes three pose changes, with a total scanning time of approximately 20 minutes per object.

After scanning, a proprietary post-processing pipeline reconstructs both the geometry and the PBR material maps. For 4K-resolution PBR material maps, the post-processing requires approximately 4 hours per object. In terms of quality, the structured-light-based shape reconstruction (Fig. 3(b)) in the post-processing achieves millimeter-level geometric accuracy. However, the material optimization process

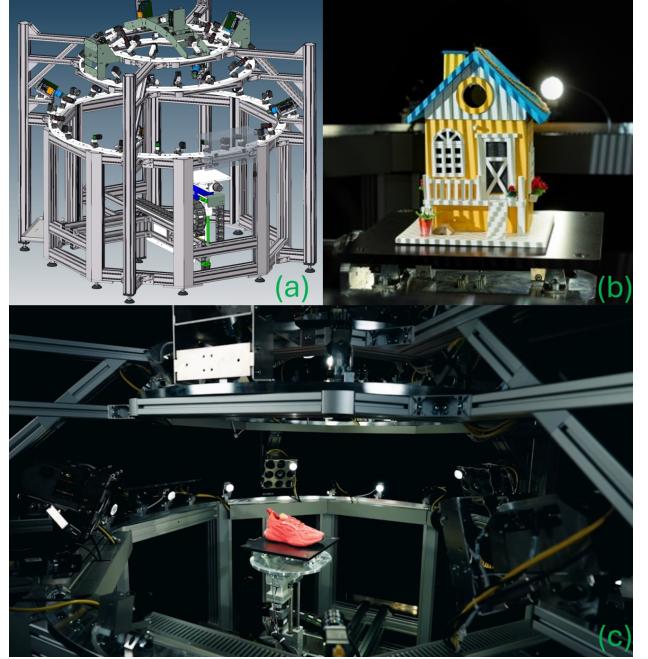


Figure 3. 3D object scanner by Covision Media®.



Figure 4. Rendered DTC models (left) v.s. Photo (Right).

performs best for diffuse objects and often struggles with glossy or shiny surfaces. To address this limitation, we hired technical artists to develop a workflow to refine materials for glossy and shiny objects, ensuring that the material quality meets the standards of a digital twin.

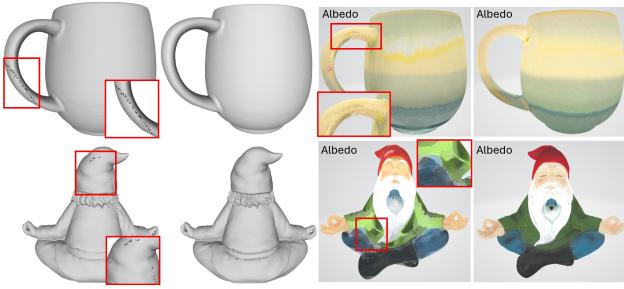


Figure 5. Shape and material (albedo) quality comparison between Stanford-ORB [39] (left) and our DTC (right).

3D Model Accuracy. To validate the material and geometry accuracy of the 3D models, we compared a rendered image of our scanned and processed model with a photograph of the same object taken inside a light box. A virtual light box was meticulously modeled to replicate the light intensity and color temperature of the real light box. The scanned object was then placed in the virtual light box to generate the rendered image. The side-by-side comparison demonstrates a remarkable match between the rendered and real images (Fig. 4).

Comparison Against Stanford-ORB. We also scanned the objects used in Stanford-ORB [39] to compare the shape and appearance quality. As illustrated in Fig. 5, the Stanford-ORB models exhibit shape artifacts and noisy, lower-quality materials compared to our models.

3.3. DSLR Evaluation Data

Within DTC, we include a DSLR-captured evaluation dataset of 50 objects from Sec. 3.2 captured under two different lighting conditions, resulting in 100 distinct image sequences. For every sequence, we provide (a) approximately 120 HDR and LDR images from different viewing directions, (b) one object pose and (c) per-image camera pose. The two lighting conditions are represented using two environment maps.

Data Capture. To ensure the DSLR evaluation data quality, we designed and built a DSLR camera rig to automate the capture process (Fig. 6). The rig is designed to rotate the cameras around the centralized object, assuming the environment lighting remains unchanged during the capture. It features a motorized rotary stage with a centrally mounted stationary platform. Attached to the rotary stage is an extrusion frame that forms the gantry arm, supported by a set of castors to bear its weight and enable smooth rotation around the central axis. The extrusion frame is equipped with adjustable camera mounts, allowing DSLR cameras to be positioned flexibly to optimize the capture setup. For our capture process, we utilized three DSLR cameras to perform a 360-degree rotation around the object, capturing images at 9-degree intervals, resulting in 120 photos per object. To ensure precise camera pose estimation, a ChArUco board



Figure 6. DSLR rig for capturing evaluation data.

was placed beneath the object during the capture. Example images from this setup are shown in Fig. 6.

Environment Maps. Following a similar approach in Stanford-ORB [39], we capture the two environment maps using chrome ball images obtained with the capture rig described earlier. With precise camera poses provided by a ChArUco board placed beneath the chrome ball, we first fit a synthetic 3D sphere to the chrome ball by optimizing its 3D position using a geometry-friendly differentiable renderer [40, 75]. Subsequently, using a differentiable Monte Carlo-based renderer [32], we refine the environment map to match the reflection on the chrome ball, employing the single-view light estimation method proposed in [78]. The coordinate system of the environment map is determined by the ChArUco detection.

Pose Registration for Camera and Object. We first obtain the initial camera poses using the ChArUco board. In most cases, for images captured from the top and middle views, the pose estimates are typically accurate. However, for bottom-view images, inaccuracies arise due to pattern distortion at grazing angles. To mitigate this issue, we refine the camera poses by fitting the rendering of a virtual ChArUco board to the real captured photo using a differentiable renderer [32, 40, 75]. Once the camera poses have been accurately refined, we optimize the object pose by minimizing the mask loss between the rendered mask and the reference one generated by [62]. By providing separate poses for camera and object anchored by ChArUco, the cameras and object are automatically aligned with the optimized environment maps.

3.4. Egocentric Evaluation Data

We include an evaluation dataset of real world recordings captured by an egocentric device paired with 100 objects from the DTC dataset. We capture the egocentric recordings using the open-sourced Project Aria device [22] and acquire the additional 3D ground truth using its machine perception tool, which includes online device calibration, device trajectory and semi-dense point clouds for each recording. Each recording contains a single 3D object and precisely aligned object poses in the Aria trajectory coordinate frame.

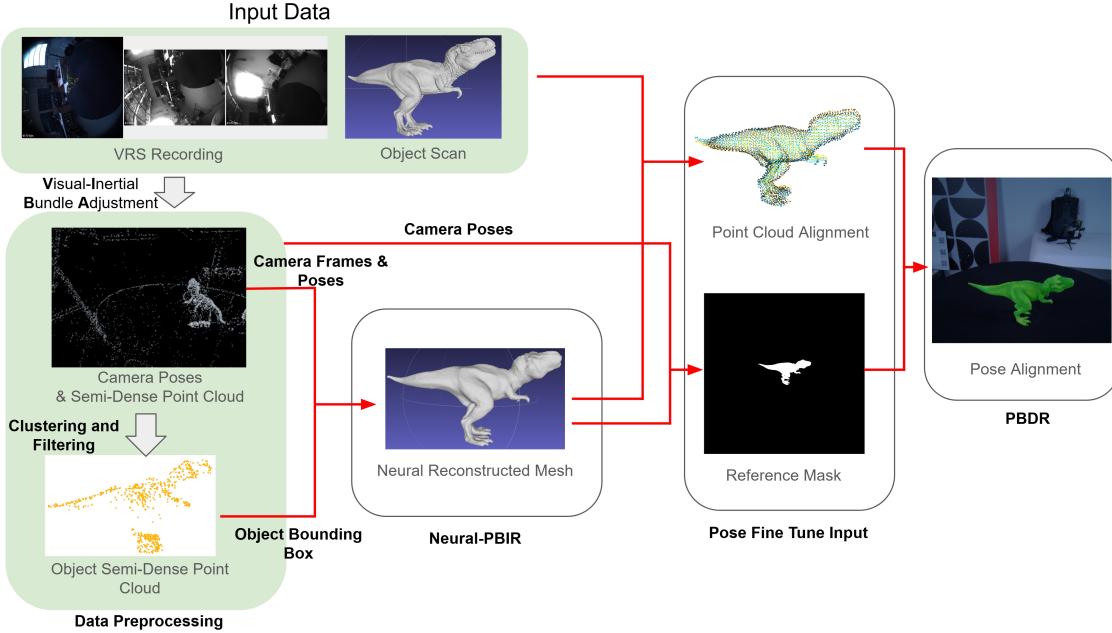


Figure 7. The workflow to align egocentric video and 3D objects. We acquire the object’s semi-dense point cloud from the egocentric recording and the neural reconstructed mesh using Neural-PBIR [62]. Finally, we align the 3D object with the rendered mask from the neural reconstructed mesh using physics-based differentiable rendering (PBDR).

Both the object and the trajectory are in metric scale. We can derive additional ground truth for each video from the aligned 3D object properties.

Data Capture. To feature real world recordings observed from human perspectives that can be representative of 3D object reconstruction for AR/MR devices or robots, we provide two types of recording trajectories for selected objects, termed *active* and *passive* respectively, collected by human wearers. The active recording features a complete 360 view of the objects, which is similar to existing object 3D novel view synthesis dataset. The *passive* recording features causal looks from human wearer, which only contain partial views of the object from certain viewing angles along the trajectory. In dataset creation, we collected the active and passive recordings in the same environment and generate their 3D information in the shared 3D space. This helps reduces potential failures when aligning the object to passive 3D recordings, which are shorter and contains less 3D information. To reduce the effect of noises and motion blur, which are common in egocentric videos in indoor low light environments, we light the capture environment to 3K+ lux illumination and used a fixed low-exposure and gain profile to collect each recording with the appropriate brightness.

Alignment between Egocentric Video and Object. We provide an illustration of the object alignment to the egocentric video in Fig. 7. Given the images, camera poses and semi-dense point cloud acquired from Project Aria tools, we employ a neural-based mesh reconstruction method [62] to

create reference meshes for generating high-quality reference masks and used that to align with the corresponding 3D object mesh. This alignment step serves as an initialization for a more precise pose refinement, which leverages differentiable rendering. For certain objects with symmetric geometry, we observe this process can introduce ambiguities in point cloud registration and subsequent failures. To address such cases, we provide a GUI to manually align and correct the object alignment. Finally, akin to the DSLR camera pose registration phase, we optimize a mask loss over object poses to achieve fine-tuned pose registration.

4. Benchmarking and Applications

We first use our DSLR and egocentric dataset as a benchmark for existing state-of-the-art methods. For inverse rendering, we design metrics to evaluate the shape and material quality of the recovered 3D object digital twin from three perspectives. For egocentric recording, we evaluate the novel-view synthesis as the initial evaluation. We include additional tasks, e.g. sparse view reconstruction in the supplementary materials for both DSLR and egocentric recordings. Finally we can demonstrate our high quality 3D digital twin models can be beneficial to robotics domain using an application in robotics manipulation.

4.1. Application to Inverse Rendering for DSLR

The DSLR dataset in DTC provides accurate ground truth, including poses, lighting and 3D models, for inverse rendering

Table 2. Benchmark comparison of existing methods on inverse rendering for DSLR. Depth SI-MSE and Shape Chamfer distance $\times 10^{-3}$.

	Geometry			Novel Scene Relighting				Novel View Synthesis			
	Depth↓	Normal↓	Shape↓	PSNR-H↑	PSNR-L↑	SSIM↑	LPIPS↓	PSNR-H↑	PSNR-L↑	SSIM↑	LPIPS↓
Neural-PIL [12]	5.71	0.25	25.02		N/A			28.42	35.76	0.882	0.096
PhySG [80]	0.31	0.16	11.31	27.28	32.86	0.959	0.049	28.54	34.46	0.964	0.045
NVDiffRec [53]	0.02	0.07	1.64	26.99	33.27	0.951	0.037	28.95	34.92	0.967	0.029
NeRD [11]	4.55	0.45	108.20	26.10	32.60	0.948	0.061	26.80	33.40	0.882	0.102
InvRender [71]	0.22	0.03	0.75	29.52	35.98	0.961	0.037	31.64	37.82	0.970	0.033
NVDiffRecMC [29]	0.02	0.06	1.34	27.78	34.55	0.952	0.042	31.27	38.17	0.972	0.032

Table 3. Benchmark on the egocentric aligned recordings.

	PSNR ↑	LPIPS ↓	SSIM ↑	Depth ↓	Normal ↓
3D-GS [37]	28.81	0.020	0.9888	0.1768	0.3301
2D-GS [31]	28.75	0.020	0.9886	0.1755	0.2112

tasks and serves as an evaluation suite to benchmark the performance of inverse rendering methods. We select six prior methods for this task and evaluate their performance using the ground truth provided by our dataset. In the following sections, we describe the data splitting strategy, evaluation metrics, and baselines.

Data Splitting. For benchmarking purposes, we select 15 objects from the DSLR dataset captured under two distinct lighting environments, resulting in a total of 30 image sequences. The selected objects encompass a diverse range of geometric and material properties to ensure a comprehensive evaluation. For each scene, 8 views are selected for testing, while the remaining views are reserved for training.

Evaluation Metrics. The metrics measure the accuracy of three aspects of baseline performance: geometry estimation, relighting, and novel view synthesis. For geometry estimation, we evaluate the accuracy of predicted depth and normal maps under held-out test views, as well as 3D meshes extracted from baseline methods, compared with the ground truth from our dataset. Relighting metrics evaluate the material decomposition quality of baselines by measuring the accuracy of predicted images under held-out lighting conditions. For view synthesis, we compare the predicted images from viewpoints unseen during training to ground truth captures. We refer to Kuang et al. [39] for metric details.

Baselines. We include the following baselines: NVDiffRec [53] and NVDiffRecMC [29], with a hybrid shape representation DMTet [60]; InvRender [85] and PhySG [80], which adopt signed distance functions (SDFs) to represent object geometry [76] and utilize implicit neural fields for material decomposition; Neural-PIL [12] and NeRD [11], which use NeRFs [52] as scene representations.

4.2. Application to Egocentric Reconstruction

Our digital twin models, aligned with real world video using the method described in Sec. 3.4, can help obtain accurate

ground truth for object-centric images that were previously difficult to acquire. We provide the first evaluation of object-centric novel view synthesis recorded from an egocentric device. We use the projected object shape given the 3D pose of the object and cameras in scene coordinates to acquire the image masks, depth and normal for each object. We selected 15 recordings from the egocentric recording sessions as the evaluation and used the *active* recordings to benchmark novel view reconstruction. For each recording, we hold out every 8th image view as a testing view.

Evaluations. We build our baselines based on the gsplats [77] implementation of the 3D Gaussian Splatting (GS) [37] and 2D GS [31], and handle the effect of lens shading from the Project Aria lens [26]. We calculate PSNR, depth and normal based on the observed objects with masks and provide SSIM and LPIPS score on images by masking out the non-object areas as black. Table 3 shows the benchmark results of the baselines. We use the same depth and normal metric in DSLR evaluation. We provide additional qualitative evaluations and analyses on egocentric data towards sparser view settings in the supplementary materials.

4.3. Application to Robotic Manipulation

High-quality object models have been leveraged in prior work to train real-world robotic agents in scenes represented explicitly [54] or implicitly [64]. These object models have also been shown to facilitate object-centric pose and lighting parameter estimation, enabling model-based planning [63]. In this section, we empirically evaluate the effectiveness of using DTC dataset objects in training robotic policies. Specifically, we consider learning robotic pushing and grasping skills in simulation.

First, we sample a subset of 24 cup category objects from the DTC dataset and 24 cup objects from Objaverse-XL [19]¹. Since not all Objaverse-XL objects come with textures, we randomize the colors of those objects uniformly in RGB space. To compute collision meshes for physical simulation, we perform convex decomposition on each object with CoACD [68]. We import these objects along with a UR5e robot equipped with a Robotiq 2F-85 (pushing) or

¹The version of Objaverse-XL used in this work excludes all 3D models sourced from Sketchfab.

Task	DTC (ours)	Objaverse-XL [19]
Pushing @ 2cm	$36.3\% \pm 1.5\%$	$25.3\% \pm 6.0\%$
Pushing @ 3cm	$43.7\% \pm 1.2\%$	$29.7\% \pm 6.0\%$
Pushing @ 5cm	$47.0\% \pm 2.6\%$	$40.3\% \pm 5.5\%$
Grasping	$42.7\% \pm 4.7\%$	$38.6\% \pm 11.0\%$

Table 4. Success rate of policies trained on data collected using objects from our DTC dataset and sampled from Objaverse-XL when evaluated on an unseen test object. Errors indicate sample standard deviation over three policy training seeds.

Robotiq 2F-140 (grasping) gripper into the PyBullet simulator [17] and collect data for each robotic task as described below. After training policies on data from each object set, we evaluate policy performance on a relatively high-quality unseen test object from the StanfordORB dataset [39].

Pushing. For the pushing task, we collect 5000 trajectories of pseudo-random robotic interaction data for each object set. For each trajectory, a single object from the considered object set and its initial position are randomly selected. Then we train a goal-conditioned neural network policy $\pi(a|o, o_g)$ where o is an image observation of the current scene and o_g is a goal image indicating the desired final object and robot position, $o, o_g \in \mathbb{R}^{256 \times 256 \times 3}$, and the action $a \in \mathbb{R}^2$ represents a change in the robot’s end-effector position in the x and y axes. The z axis end-effector height is held fixed. We sample goals for training via hindsight relabeling [4, 20, 24]. We then perform evaluation on 100 randomly sampled test goals manipulating an unseen test cup from StanfordORB [39].

Grasping. For grasping, we collect 5000 successful grasp examples for each object set by first placing a single object into the scene, randomizing the object identity and initial position. We then randomly sample candidate grasp poses in a radius around the object’s position and simulating their outcomes, rejecting unsuccessful grasps. We train a grasping policy $\pi(a|o)$ where $o \in \mathbb{R}^{256 \times 256 \times 3}$ is an image observation of the scene and $a \in \mathbb{R}^4$ represents the x, y, z position and θ yaw rotation of the robot end-effector at which to attempt the grasp. Again we use 100 test cup object poses.

Results. We report the results in Table 4 and Fig. 8. We find that across both pushing and grasping tasks, policies trained on DTC dataset objects outperform those trained on Objaverse-XL objects when evaluated on the unseen test object. For pushing, we report performance by defining binary success thresholds based on the final Euclidean distance of the object position to the goal position. Training on DTC objects appears to be especially helpful at enabling policies to make finer adjustments, improving pushing success rates at stricter thresholds. Additional experimental details can be found in the supplementary.

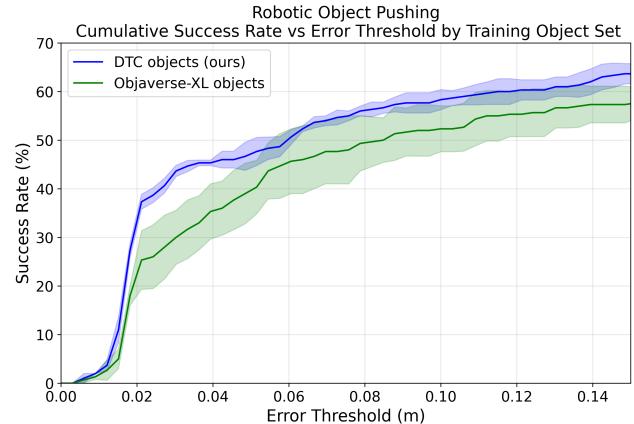


Figure 8. Success rates on robotic pushing task when training with our DTC objects and sampled Objaverse-XL objects. Particularly at lower error thresholds, policies trained on DTC objects outperform those trained using Objaverse-XL objects. Shaded bars represent sample standard deviations over policy training random seeds.

5. Conclusion

We presented a new large scale photorealistic 3D digital twin dataset with the real world recordings that contain its real world counterpart. We provide extensive evaluations of baselines on our DTC dataset serving as new benchmark for inverse rendering and novel view synthesis task. We also demonstrated that high quality digital twin models can be beneficial to applications in robotics domain. We believe our efforts can empower the research community to build and leverage digital twin models for future applications.

Limitations. Achieving high quality digital twin models currently requires deliberate hardware setup and human efforts in refinement. Solving this challenge without sacrificing quality can significantly further enhance the volume of digital twin models. Our hardware is also limited to objects within a certain size and can not yet recover objects that are deformable, highly specular, or transparent.

Future work. The existing digital twin model creation in DTC dataset involves lengthy post-processing and may require subjective human refinement, hindering the automation of model generation. However, recent advancements in physics-based differentiable rendering hold promise for enabling faster and more accurate creation of digital twins, especially for material reconstruction. Furthermore, building large-scale digital twins for applications will necessitate efforts to enhance the diversity in object appearance (e.g., transparent objects) and to capture additional attributes, such as physical properties and functionalities.

For robotics applications, while deploying manipulation policies learned in simulation to the real world remains generally challenging, we hope that the high-quality digital twin data provided by DTC can serve as a stepping stone towards effective sim-to-real transfer.

Acknowledgments. This work is in part supported by NSF CCRI #2120095, RI #2211258, and RI #2338203. YZ is in part supported by the Stanford Interdisciplinary Graduate Fellowship.

References

- [1] Covision 3D object scanner. <https://www.covisionmedia.ai/en/scanners>. 4
- [2] Kosmas Alexopoulos, Nikolaos Nikolakis, and George Chrysoulouris. Digital twin-driven supervised machine learning for the development of artificial intelligence applications in manufacturing. *International Journal of Computer Integrated Manufacturing*, 33(5):429–439, 2020. 2
- [3] Moayad Aloqaily, Ouns Bouachir, Fakhri Karray, Ismael Al Ridhawi, and Abdulmotaleb El Saddik. Integrating digital twin and advanced intelligent technologies to realize the metaverse. *IEEE Consumer Electronics Magazine*, 12(6):47–55, 2022. 2
- [4] Marcin Andrychowicz, Filip Wolski, Alex Ray, Jonas Schneider, Rachel Fong, Peter Welinder, Bob McGrew, Josh Tobin, OpenAI Pieter Abbeel, and Wojciech Zaremba. Hindsight experience replay. In *NeurIPS*, 2017. 8
- [5] Jonathan T Barron and Jitendra Malik. Shape, illumination, and reflectance from shading. *IEEE TPAMI*, 37(8):1670–1687, 2014. 3
- [6] Sean Bell, Kavita Bala, and Noah Snavely. Intrinsic images in the wild. *ACM TOG*, 33(4):1–12, 2014. 3
- [7] Sai Bi, Zexiang Xu, Kalyan Sunkavalli, Miloš Hašan, Yannick Hold-Geoffroy, David Kriegman, and Ravi Ramamoorthi. Deep reflectance volumes: Relightable reconstructions from multi-view photometric images. In *ECCV*, 2020. 3
- [8] Sai Bi, Zexiang Xu, Kalyan Sunkavalli, David Kriegman, and Ravi Ramamoorthi. Deep 3D capture: Geometry and reflectance from sparse multi-view images. In *CVPR*, 2020. 4
- [9] Martin Bichsel and Alex P Pentland. A simple algorithm for shape from shading. In *CVPR*, 1992. 3
- [10] Mark Boss, Varun Jampani, Kihwan Kim, Hendrik P.A. Lensch, and Jan Kautz. Two-shot Spatially-varying BRDF and Shape Estimation. In *CVPR*, 2020. 4
- [11] Mark Boss, Raphael Braun, Varun Jampani, Jonathan T. Barron, Ce Liu, and Hendrik P.A. Lensch. NeRD: Neural reflectance decomposition from image collections. In *ICCV*, 2021. 3, 7
- [12] Mark Boss, Varun Jampani, Raphael Braun, Ce Liu, Jonathan Barron, and Hendrik Lensch. Neural-PIL: Neural pre-integrated lighting for reflectance decomposition. In *NeurIPS*, 2021. 7
- [13] Mark Boss, Andreas Engelhardt, Abhishek Kar, Yuanzhen Li, Deqing Sun, Jonathan Barron, Hendrik Lensch, and Varun Jampani. SAMURAI: Shape and material from unconstrained real-world arbitrary image collections. In *NeurIPS*, 2022. 3
- [14] Guangyan Cai, Kai Yan, Zhao Dong, Ioannis Gkioulekas, and Shuang Zhao. Physics-based inverse rendering using combined implicit and explicit geometries. *CGF*, 41(4):129–138, 2022. 3
- [15] Wenzheng Chen, Joey Litalien, Jun Gao, Zian Wang, Clement Fuji Tsang, Sameh Khamis, Or Litany, and Sanja Fidler. DIB-R++: learning to predict lighting and material with a hybrid differentiable renderer. In *NeurIPS*, 2021. 3
- [16] Jasmine Collins, Shubham Goel, Kenan Deng, Achleshwar Luthra, Leon Xu, Erhan Gundogdu, Xi Zhang, Tomas F Yago Vicente, Thomas Dideriksen, Himanshu Arora, et al. ABO: Dataset and benchmarks for real-world 3D object understanding. In *CVPR*, 2022. 2, 3
- [17] Erwin Coumans and Yunfei Bai. Pybullet, a python module for physics simulation for games, robotics and machine learning. <http://pybullet.org>, 2016–2019. 8
- [18] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3D objects. In *CVPR*, 2023. 2, 3
- [19] Matt Deitke, Ruoshi Liu, Matthew Wallingford, Huong Ngo, Oscar Michel, Aditya Kusupati, Alan Fan, Christian Laforte, Vikram Voleti, Samir Yitzhak Gadre, et al. Objaverse-XL: A universe of 10m+ 3D objects. *NeurIPS*, 2024. 3, 7, 8
- [20] Yiming Ding, Carlos Florensa, Pieter Abbeel, and Mariano Phielipp. Goal-conditioned imitation learning. In *NeurIPS*, 2019. 8
- [21] Laura Downs, Anthony Francis, Nate Koenig, Brandon Kinman, Ryan Hickman, Krista Reymann, Thomas B McHugh, and Vincent Vanhoucke. Google scanned objects: A high-quality dataset of 3D scanned household items. In *ICRA*, 2022. 2, 3
- [22] Jakob Engel, Kiran Somasundaram, Michael Goesele, Albert Sun, Alexander Gamino, Andrew Turner, Arjang Talat-tor, Arnie Yuan, Bilal Souti, Brighid Meredith, Cheng Peng, Chris Sweeney, Cole Wilson, Dan Barnes, Daniel DeTone, David Caruso, Derek Valleroy, Dinesh Ginjupalli, Duncan Frost, Edward Miller, Elias Muegller, Evgeniy Oleinik, Fan Zhang, Guruprasad Somasundaram, Gustavo Solaira, Harry Lanaras, Henry Howard-Jenkins, Huixuan Tang, Hyo Jin Kim, Jaime Rivera, Ji Luo, Jing Dong, Julian Straub, Kevin Bailey, Kevin Eckenhoff, Lingni Ma, Luis Pesqueira, Mark Schwesinger, Maurizio Monge, Nan Yang, Nick Charron, Nikhil Raina, Omkar Parkhi, Peter Borschowa, Pierre Moulon, Prince Gupta, Raul Mur-Artal, Robbie Pennington, Sachin Kulkarni, Sagar Miglani, Santosh Gondi, Saransh Solanki, Sean Diener, Shangyi Cheng, Simon Green, Steve Saarinen, Suvam Patra, Tassos Mourikis, Thomas Whelan, Tripti Singh, Vasileios Balntas, Vijay Baiyya, Wilson Dreewes, Xiaqing Pan, Yang Lou, Yipu Zhao, Yusuf Mansour, Yuyang Zou, Zhaoyang Lv, Zijian Wang, Mingfei Yan, Carl Ren, Renzo De Nardi, and Richard Newcombe. Project Aria: A new tool for egocentric multi-modal ai research, 2023. 5
- [23] Ruohan Gao, Yen-Yu Chang, Shivani Mall, Li Fei-Fei, and Jiajun Wu. ObjectFolder: A dataset of objects with implicit visual, auditory, and tactile representations. In *CoRL*, 2021. 2
- [24] Dibya Ghosh, Abhishek Gupta, Ashwin Reddy, Justin Fu, Coline Devin, Benjamin Eysenbach, and Sergey Levine. Learning to reach goals via iterated supervised learning. In *ICML*, 2019. 8
- [25] Roger Grosse, Micah K Johnson, Edward H Adelson, and

- William T Freeman. Ground truth dataset and baseline evaluations for intrinsic image algorithms. In *ICCV*, 2009. 3
- [26] Qiao Gu, Zhaoyang Lv, Duncan Frost, Simon Green, Julian Straub, and Chris Sweeney. Egolifter: Open-world 3d segmentation for egocentric perception. In *ECCV*, pages 382–400. Springer, 2024. 7
- [27] Agrim Gupta, Piotr Dollar, and Ross Girshick. LVIS: A dataset for large vocabulary instance segmentation. In *CVPR*, 2019. 4
- [28] Wassim Hamidouche, Lina Bariah, and Mérouane Debbah. Immersive media and massive twinning: Advancing toward the metaverse. *IEEE Communications Magazine*, 62(7):20–32, 2024. 2
- [29] Jon Hasselgren, Nikolai Hofmann, and Jacob Munkberg. Shape, light & material decomposition from images using Monte Carlo rendering and denoising. In *NeurIPS*, 2022. 4, 7
- [30] Yicong Hong, Kai Zhang, Jiuxiang Gu, Sai Bi, Yang Zhou, Difan Liu, Feng Liu, Kalyan Sunkavalli, Trung Bui, and Hao Tan. LRM: Large reconstruction model for single image to 3D. In *The Twelfth International Conference on Learning Representations*, 2024. 2
- [31] Binbin Huang, Zehao Yu, Anpei Chen, Andreas Geiger, and Shenghua Gao. 2D gaussian splatting for geometrically accurate radiance fields. In *SIGGRAPH 2024 Conference Papers*, 2024. 3, 7
- [32] Wenzel Jakob, Sébastien Speierer, Nicolas Roussel, Merlin Nimier-David, Delio Vicini, Tizian Zeltner, Baptiste Nicolet, Miguel Crespo, Vincent Leroy, and Ziyi Zhang. Mitsuba 3 renderer, 2022. <https://mitsuba-renderer.org>. 5
- [33] Michael Janner, Jiajun Wu, Tejas Kulkarni, Ilker Yildirim, and Joshua B Tenenbaum. Self-supervised intrinsic image decomposition. In *NeurIPS*, 2017. 4
- [34] Rasmus Jensen, Anders Dahl, George Vogiatzis, Engin Tola, and Henrik Aanæs. Large scale multi-view stereopsis evaluation. In *CVPR*, 2014. 2, 3
- [35] Haian Jin, Isabella Liu, Peijia Xu, Xiaoshuai Zhang, Songfang Han, Sai Bi, Xiaowei Zhou, Zexiang Xu, and Hao Su. TensoIR: Tensorial inverse rendering. In *CVPR*, 2023. 2
- [36] James T Kajiya. The rendering equation. In *Proceedings of the 13th annual conference on Computer graphics and interactive techniques*, pages 143–150, 1986. 3
- [37] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3D gaussian splatting for real-time radiance field rendering. *ACM TOG*, 2023. 2, 3, 7
- [38] Zhengfei Kuang, Kyle Olszewski, Menglei Chai, Zeng Huang, Panos Achlioptas, and Sergey Tulyakov. NeROIC: neural rendering of objects from online image collections. *ACM TOG*, 41(4):1–12, 2022. 3
- [39] 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. In *Advances in Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. 3, 5, 7, 8
- [40] Samuli Laine, Janne Hellsten, Tero Karras, Yeongho Seol, Jaakko Lehtinen, and Timo Aila. Modular primitives for high-performance differentiable rendering. *ACM Transactions on Graphics*, 39(6), 2020. 5
- [41] Jiahao Li, Hao Tan, Kai Zhang, Zexiang Xu, Fujun Luan, Yinghao Xu, Yicong Hong, Kalyan Sunkavalli, Greg Shakhnarovich, and Sai Bi. Instant3D: Fast text-to-3D with sparse-view generation and large reconstruction model. In *ICLR*, 2024. 2
- [42] Min Li, Zhenglong Zhou, Zhe Wu, Boxin Shi, Changyu Diao, and Ping Tan. Multi-view photometric stereo: A robust solution and benchmark dataset for spatially varying isotropic materials. *IEEE TIP*, 29:4159–4173, 2020. 3
- [43] Xuanlin Li, Kyle Hsu, Jiayuan Gu, Karl Pertsch, Oier Mees, Homer Rich Walke, Chuyuan Fu, Ishikaa Lunawat, Isabel Sieh, Sean Kirmani, Sergey Levine, Jiajun Wu, Chelsea Finn, Hao Su, Quan Vuong, and Ted Xiao. Evaluating real-world robot manipulation policies in simulation. *arXiv preprint arXiv:2405.05941*, 2024. 2
- [44] Zhengqin Li, Kalyan Sunkavalli, and Manmohan Chandraker. Materials for masses: SVBRDF acquisition with a single mobile phone image. In *ECCV*, 2018. 3
- [45] Zhengqin Li, Zexiang Xu, Ravi Ramamoorthi, Kalyan Sunkavalli, and Manmohan Chandraker. Learning to reconstruct shape and spatially-varying reflectance from a single image. *ACM TOG*, 37(6):1–11, 2018. 3, 4
- [46] Isabella Liu, Linghao Chen, Ziyang Fu, Liwen Wu, Haian Jin, Zhong Li, Chin Ming Ryan Wong, Yi Xu, Ravi Ramamoorthi, Zexiang Xu, and Hao Su. OpenIllumination: A multi-illumination dataset for inverse rendering evaluation on real objects. *NeurIPS*, 2023. 2, 3
- [47] Stephen Lombardi and Ko Nishino. Reflectance and natural illumination from a single image. In *ECCV*, 2012. 3
- [48] Stephen Lombardi and Ko Nishino. Reflectance and illumination recovery in the wild. *IEEE TPAMI*, 2015. 3
- [49] Fujun Luan, Shuang Zhao, Kavita Bala, and Zhao Dong. Unified shape and SVBRDF recovery using differentiable Monte Carlo rendering. *CGF*, 40(4):101–113, 2021. 3
- [50] Stephen Robert Marschner. *Inverse Rendering for Computer Graphics*. Cornell University, 1998. 3
- [51] Nelson Max. Optical models for direct volume rendering. *IEEE TVCG*, 1(2):99–108, 1995. 3
- [52] 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. In *ECCV*, 2020. 2, 3, 7
- [53] 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 *CVPR*, 2022. 4, 7
- [54] Soroush Nasiriany, Abhiram Maddukuri, Lance Zhang, Adeet Parikh, Aaron Lo, Abhishek Joshi, Ajay Mandlekar, and Yuke Zhu. Robocasa: Large-scale simulation of everyday tasks for generalist robots. In *Robotics: Science and Systems (RSS)*, 2024. 7
- [55] Michael Niemeyer, Lars Mescheder, Michael Oechsle, and Andreas Geiger. Differentiable volumetric rendering: Learning implicit 3D representations without 3D supervision. In *CVPR*, 2020. 4
- [56] Geoffrey Oxholm and Ko Nishino. Multiview shape and reflectance from natural illumination. In *CVPR*, 2014. 3

- [57] Geoffrey Oxholm and Ko Nishino. Shape and reflectance estimation in the wild. *IEEE TPAMI*, 2015. 3
- [58] Xiaqing Pan, Nicholas Charron, Yongqian Yang, Scott Peters, Thomas Whelan, Chen Kong, Omkar Parkhi, Richard Newcombe, and Yuheng (Carl) Ren. Aria digital twin: A new benchmark dataset for egocentric 3D machine perception. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 20133–20143, 2023. 3
- [59] Viktor Rudnev, Mohamed Elgharib, William Smith, Lingjie Liu, Vladislav Golyanik, and Christian Theobalt. NeRF for outdoor scene relighting. In *ECCV*, pages 615–631. Springer, 2022. 3
- [60] Tianchang Shen, Jun Gao, Kangxue Yin, Ming-Yu Liu, and Sanja Fidler. Deep marching tetrahedra: a hybrid representation for high-resolution 3D shape synthesis. In *NeurIPS*, 2021. 7
- [61] Jian Shi, Yue Dong, Hao Su, and Stella X Yu. Learning non-Lambertian object intrinsics across ShapeNet categories. In *CVPR*, 2017. 2, 3, 4
- [62] Cheng Sun, Guangyan Cai, Zhengqin Li, Kai Yan, Cheng Zhang, Carl Marshall, Jia-Bin Huang, Shuang Zhao, and Zhao Dong. Neural-PBIR reconstruction of shape, material, and illumination. In *ICCV*, 2023. 2, 4, 5, 6
- [63] Stephen Tian, Yancheng Cai, Hong-Xing Yu, Sergey Zakharov, Katherine Liu, Adrien Gaidon, Yunzhu Li, and Jiajun Wu. Multi-object manipulation via object-centric neural scattering functions. In *CVPR*, 2023. 7
- [64] Marcel Torne, Anthony Simeonov, Zechu Li, April Chan, Tao Chen, Abhishek Gupta, and Pulkit Agrawal. Reconciling reality through simulation: A real-to-sim-to-real approach for robust manipulation. *Arxiv*, 2024. 7
- [65] Marco Toschi, Riccardo De Matteo, Riccardo Spezialetti, Daniele De Gregorio, Luigi Di Stefano, and Samuele Salti. Relight my NeRF: A dataset for novel view synthesis and relighting of real world objects. In *CVPR*, 2023. 3
- [66] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. NeuS: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. In *NeurIPS*, 2021. 4
- [67] Henrique Weber, Donald Prévost, and Jean-François Lalonde. Learning to estimate indoor lighting from 3D objects. In *3DV*, 2018. 3
- [68] Xinyue Wei, Minghua Liu, Zhan Ling, and Hao Su. Approximate convex decomposition for 3D meshes with collision-aware concavity and tree search. *ACM TOG*, 2022. 7
- [69] Xinyue Wei, Kai Zhang, Sai Bi, Hao Tan, Fujun Luan, Valentin Deschaintre, Kalyan Sunkavalli, Hao Su, and Zexiang Xu. MeshLRM: Large reconstruction model for high-quality mesh. *arXiv preprint arXiv:2404.12385*, 2024. 2
- [70] Felix Wimbauer, Shangzhe Wu, and Christian Rupprecht. De-rendering 3D objects in the wild. In *CVPR*, 2022. 3, 4
- [71] Haoqian Wu, Zhipeng Hu, Lincheng Li, Yongqiang Zhang, Changjie Fan, and Xin Yu. NeFII: Inverse rendering for reflectance decomposition with near-field indirect illumination. In *CVPR*, 2023. 4, 7
- [72] Shangzhe Wu, Ameesh Makadia, Jiajun Wu, Noah Snavely, Richard Tucker, and Angjoo Kanazawa. De-rendering the world’s revolutionary artefacts. In *CVPR*, 2021. 3, 4
- [73] Tong Wu, Jiarui Zhang, Xiao Fu, Yuxin Wang, Liang Pan, Jiawei Ren, Wayne Wu, Lei Yang, Jiaqi Wang, Chen Qian, Dahua Lin, and Ziwei Liu. OmniObject3D: Large-vocabulary 3D object dataset for realistic perception, reconstruction and generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2, 3
- [74] Kohei Yamashita, Yuto Enyo, Shohei Nobuhara, and Ko Nishino. nLMVS-Net: Deep non-lambertian multi-view stereo. In *WACV*, 2023. 3
- [75] Kai Yan and Shuang Zhao. PSDR-JIT: Path-space differentiable renderer, 2020. <https://github.com/andyyankai/psdr-jit>. 5
- [76] Lior Yariv, Yoni Kasten, Dror Moran, Meirav Galun, Matan Atzmon, Basri Ronen, and Yaron Lipman. Multiview neural surface reconstruction by disentangling geometry and appearance. In *NeurIPS*, 2020. 7
- [77] Vickie Ye, Ruilong Li, Justin Kerr, Matias Turkulainen, Brent Yi, Zhuoyang Pan, Otto Seiskari, Jianbo Ye, Jeffrey Hu, Matthew Tancik, and Angjoo Kanazawa. gsplat: An open-source library for Gaussian splatting. *arXiv preprint arXiv:2409.06765*, 2024. 7
- [78] Hong-Xing Yu, Samir Agarwala, Charles Herrmann, Richard Szeliski, Noah Snavely, Jiajun Wu, and Deqing Sun. Accidental light probes. In *CVPR*, 2023. 3, 5
- [79] Hong-Xing Yu, Michelle Guo, Alireza Fathi, Yen-Yu Chang, Eric Ryan Chan, Ruohan Gao, Thomas Funkhouser, and Jiajun Wu. Learning object-centric neural scattering functions for free-viewpoint relighting and scene composition. *TMLR*, 2023. 3
- [80] 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 *CVPR*, 2021. 4, 7
- [81] Kai Zhang, Fujun Luan, Zhengqi Li, and Noah Snavely. IRON: Inverse rendering by optimizing neural SDFs and materials from photometric images. In *CVPR*, 2022. 4
- [82] Kai Zhang, Sai Bi, Hao Tan, Yuanbo Xiangli, Nanxuan Zhao, Kalyan Sunkavalli, and Zexiang Xu. GS-LRM: Large reconstruction model for 3D gaussian splatting. *arXiv preprint arXiv:2404.19702*, 2024. 2
- [83] Ruo Zhang, Ping-Sing Tsai, James Edwin Cryer, and Mubarak Shah. Shape-from-shading: a survey. *IEEE TPAMI*, 21(8): 690–706, 1999. 3
- [84] Xiuming Zhang, Pratul P Srinivasan, Boyang Deng, Paul Debevec, William T Freeman, and Jonathan T Barron. NeR-Factor: Neural factorization of shape and reflectance under an unknown illumination. *ACM TOG*, 40(6):1–18, 2021. 3
- [85] Yuanqing Zhang, Jiaming Sun, Xingyi He, Huan Fu, Rongfei Jia, and Xiaowei Zhou. Modeling indirect illumination for inverse rendering. In *CVPR*, 2022. 4, 7
- [86] Yunzhi Zhang, Shangzhe Wu, Noah Snavely, and Jiajun Wu. Seeing a rose in five thousand ways. In *CVPR*, 2023. 4