Breaking the 3D Dataset Bottleneck: Fast Scalable Generation of Aligned 3D Assets from Scratch for Category 6D Pose Estimation and Robotic Grasping

Duret Guillaume[1,3], Danylo Mazurak[1], Florence Zara[2], Jan Peters[3], Liming Chen[1]
[1]Centrale Lyon, CNRS, LIRIS, UMR5205, F-69130 Ecully, France
[2]UCBL, CNRS, LIRIS, UMR5205, F-69622 Villeurbanne, France
[3]Intelligent Autonomous Systems Lab, Technical University of Darmstadt, 64289 Darmstadt, Germany

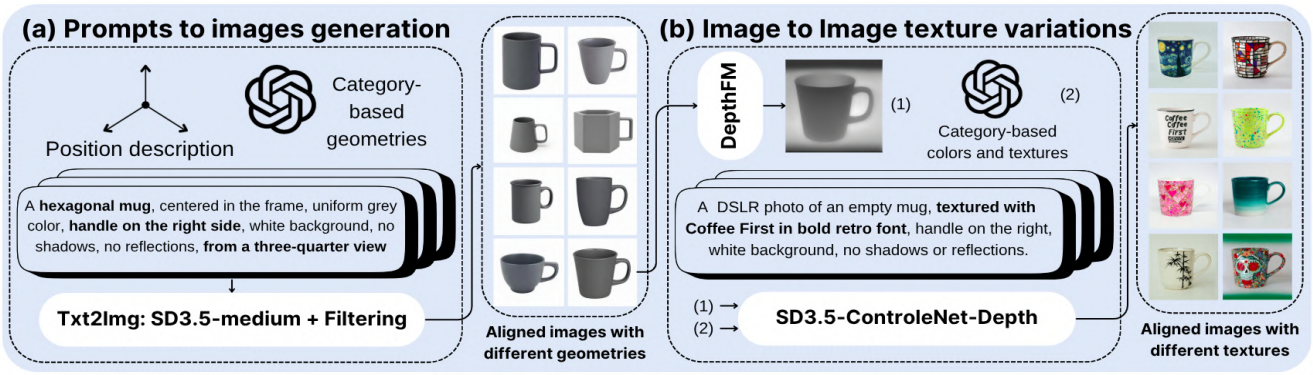

Figure 1. Our text-to-image pipeline: (a) Category-based geometry prompt engineering and images generation; (b) Depth-conditioned image generation for texture variation and automatic alignment.

Abstract

While 2D vision has been revolutionized by large-scale datasets like ImageNet, 3D vision remains constrained by the scarcity of high-quality, canonically aligned data. We introduce the first scalable, automated framework that generates complete category-level 6D pose datasets directly from text prompts, bypassing the need for existing 3D assets. Our method overcomes key challenges by: (1) ensuring reliable, scalable asset generation via a controlled text-to-image-to-3D pipeline; (2) enforcing built-in canonical alignment through depth-conditioned generation, achieving a 96% pose consistency rate; and (3) enabling large-scale 6D annotation via mixed reality rendering. The pipeline produces high-quality, aligned 3D meshes in under 3 minutes per object—a 5–20× speedup over traditional scanning. We generate over 1,000 instances for each of the 153 categories in the Omni6Dpose benchmark, culminating in 153,000 aligned meshes—a $> 40\times$ increase in instances per category over previous aligned real-world datasets. Extensive evaluation demonstrates competitive zero-shot sim2real transfer on the NOCS 6D pose benchmark and superior robotic grasping performance in both simulation and real-world zero-shot transfer, where aligned meshes prove essential for success. We release the largest publicly available aligned 3D mesh dataset, largest category-level 6D pose dataset, grasping simulation environments, and open-source pipeline, providing a critical step toward foundation models for 3D understanding and enabling efficient, unlimited generation of task-specific 3D data from scratch. The code and datasets can be found at https://genomni3d.github.io/

1. Introduction

The field of 2D computer vision has been revolutionized by foundation models trained on massive-scale datasets [23–25]. In contrast, 3D vision advancement remains constrained by a fundamental limitation: the scarcity of high-quality, diverse, and scalable annotated 3D data. This bottleneck is critical given that advanced 3D understanding underpins numerous applications including robotics, augmented reality, and autonomous systems. Unlike 2D data, effective 3D learning requires not just annotations but also consistent category-level alignment across instances.

Category-level 6D pose estimation—predicting an ob-

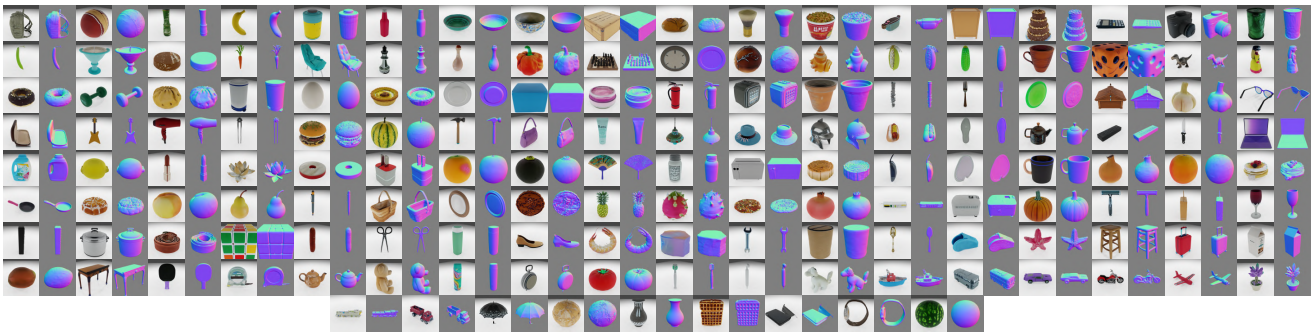

Figure 2. GenOmni3D Dataset: RGB and Normal images for objects from all 153 categories.

ject's 3D position and orientation from a single RGB-D image without instance-specific models—epitomizes this challenge. While this formulation reduces reliance on exact object geometry, it demands extensive category-specific datasets with aligned 3D assets exhibiting substantial intra-class shape and texture variation. The creation of such datasets faces three fundamental bottlenecks:

1. Asset Collection: Existing approaches depend on labor-intensive 3D scanning (15–60 minutes per object) or limited pre-existing 3D repositories. Current datasets fall into three problematic categories: (1) synthetic collections lacking realism [3]; (2) high-quality scanned datasets with limited scale [32]; and (3) large-scale, internet-sourced repositories suffering from inconsistent mesh quality, poor alignment, and sparse category coverage [7]. This scarcity is especially pronounced for category-level 6D pose estimation, where existing datasets cover limited categories with insufficient instance diversity for robust learning.

2. Mesh Alignment: Canonical alignment across object instances is crucial for effective category-based 3D understanding. However, achieving consistent precise alignment requires extensive manual effort, fundamentally hindering large-scale dataset creation. While existing datasets provide some aligned assets, the availability of large-scale, consistently aligned datasets remains largely unmet in the 6D pose estimation domain, forcing methods to rely on inter-category transfer rather than strong per-category learning.

3. Pose Annotation: Real-world 6D pose annotation typically involves pre-scanned assets and iterative optimization (e.g., ICP), making the process error-prone and challenging to scale across categories and scenes. Synthetic datasets offer partial solutions but are limited in generating diverse, realistic environments, restricting their applicability for sim-to-real transfer.

For these challenges, we present an end-to-end scalable pipeline that transforms text prompts into complete 6D pose and grasping datasets. Our contributions encompass:

Generative Pipeline for Aligned 3D Assets: We introduce the first automated framework that converts category descriptions into aligned 3D meshes with 96% pose consistency—dramatically improving upon the 57% achieved by prior work [10]. The pipeline produces production-ready assets in under 3 minutes per object, achieving 5–20× speedup over traditional scanning.

Benchmarking Dataset Generation: We reproduce, benchmark, and open-source state-of-the-art category-level 6D dataset generation pipelines, including full 3D simulation from Omni6D [38] and mixed-reality rendering from Omni6Dpose [36]. We generate comprehensive evaluation datasets demonstrating superior category-level 6D pose estimation and sim2real transfer on the NOCS benchmark.

Large-Scale Dataset Releases: Leveraging our pipeline, we release two datasets: (1) the largest aligned 3D mesh dataset (153K meshes across 153 categories) and (2) the largest category-level 6D pose dataset (1.2M images with full annotations), providing unprecedented scale for 3D foundation model development and enabling superior benchmarking across the Omni6D benchmark [38].

Grasping Simulation and Real-World Validation: We develop complete custom grasping simulation environments from scratch in SAPIEN [34], achieving 87.8% grasp success rate and significant improvements in shape completion (0.475 IoU vs 0.314), and demonstrate superior real-world zero-shot transfer. Our validation shows aligned meshes are essential for optimal robotic manipulation performance.

By decoupling dataset creation from manual processes, our work establishes an efficient framework for generating complete 3D understanding ecosystems from scratch. The released pipeline and datasets enable the community to overcome the critical data scarcity that has long hindered progress in 3D perception and manipulation, paving the way for scalable foundation models in 3D vision.

2. Related Work

2.1. 3D Datasets

As shown in Table 1, existing 3D datasets can be categorized into four main types: synthetic datasets such as ShapeNet [3] (55K objects) established early bench-

marks for 3D object recognition, while subsequent datasets including ModelNet [33], 3D-FUTURE [12], ABO [5], and Toy4K [26] improved category variety but generally lacked photorealism; real-world scan datasets including GSO (1K objects), ABO (2K articulated objects), and OmniObject3D [32] (6K objects) offer high-fidelity scans but face scalability limitations due to time-consuming acquisition processes requiring 15-60 minutes per object; large-scale internet collections such as Objaverse1.0 [6] (800K+ meshes) and ObjaverseXL [7] (10.2M meshes) provide unprecedented scale but suffer from inconsistent mesh quality and sparse category coverage, with only 127 mugs available in Objaverse1.0 as one example of this limitation; and most similar to our work, generated collections include GenVegeFruits3D [10] which generates 3D assets but is limited to symmetric produce, avoiding the challenges posed by arbitrary shapes, with its CPU-based texturing pipeline introducing significant computational overhead as discussed in Section 3.2 and Table 3.

Table 1. Comparison of existing 3D mesh datasets, highlighting their number of instances by categories. R/S/SAI indicates whether the dataset consists of real-world scanned objects (R), synthetic assets created by artists (S), or assets from generative 3D models (SAI).

3D dataset	R/S/SAI	#Obj	#Cat	#O/C	Ali	Quality	Time
ShapeNet [3]	S	51k	55	927	**Y**	*	N/A
ModelNet [33]	S	12k	40	300	**Y**	*	N/A
3D-Future [12]	S	16k	34	470	3D-FutureY	*	N/A
ABO [5]	S	8k	63	159	**Y**	*	N/A
Toy4K [26]	R	4k	105	38	**Y**	*	15m-1h
GSO [9]	R	1k	17	59	**Y**	***	15m-1h
AKB-48 [20]	R	2k	48	42	**Y**	***	15m-1h
OmniObject3D [32]	R	6k	190	32	Y	***	15m-1h
Objaverse1.0 [6]	R+S	800k	-	100	**N**	**	N/A
ObjaverseXL [7]	R+S	**10.2M**	-	-	N	*	N/A
GenVegeFruits3D [10]	SAI	100K	**100**	**1000**	**Y**	***	15min
GenNocs3D (Ours)	SAI	6K	**6**	**1000**	**Y**	***	**3min**
GenOmni3D (Ours)	SAI	153K	**153**	**1000**	**Y**	***	**3min**

As summarized in Table 1, our datasets, *GenNOCS3D* and *GenOmni3D*, uniquely combine large-scale instance diversity with built-in canonical alignment while achieving significantly faster generation times. This combination enables the efficient creation of high-quality, consistently aligned meshes that are directly usable for downstream 3D understanding tasks, without dependency on limited existing 3D datasets.

2.2. Category 6D Pose Datasets

The scarcity of high-quality 6D pose annotations remains a major bottleneck for category-level pose estimation. Real-world annotation is costly, while synthetic datasets require diverse 3D assets—many of which need extensive preprocessing, limiting scalability. NOCS [29] (6 categories) established the first benchmark, followed by PhoCAL [31], Wild6D [13], and HouseCat6D [18], which improved diversity but still suffer from sparse category coverage. Recent efforts like Omni6D [38] and Omni6Dpose [36] combine synthetic and real 3D scanned data but face two key issues: (1) limited instance diversity due to reliance on OmniObject3D [32] scans, and (2) reproducibility challenges from closed-source rendering pipelines [36]. Indeed, these benchmarks have very few mesh instances per category and rely on inter-class transfer learning rather than strong 3D learning of each category. This highlights the need for open, scalable frameworks that reduce dependency on pre-scanned assets while ensuring diversity and reproducibility.

In contrast, our approach enables the generation of unlimited, canonically aligned 3D assets across arbitrary categories. We showcase the proposed approach through *GenNOCS6D* and *GenOmni6D*, which integrate complementary rendering pipelines for sim2real transfer. We further release both pipelines along with large-scale datasets comprising 600K and 1.2M high-quality synthetic images respectively, to support scalable training and reproducibility in category-level 6D pose estimation.

3. From category prompts to high-quality textured 3D mesh generation

This section presents our pipeline for generating high-quality, textured 3D meshes from only a category input. Section 3.1 describes the text-to-image-to-3D architecture, Section 3.2 examines the role of depth conditioning, and Section 3.3 benchmarks 3D reconstruction methods for quality and efficiency. The pipeline enables fully automated generation of 1,000 aligned meshes per category from just 100 depth images, reducing manual filtering by over 15× compared to prior work [10].

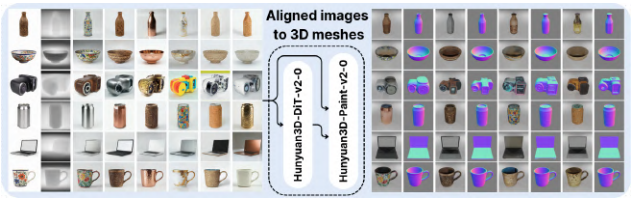

Figure 3. Qualitative examples of final textured images of the 6 categories in the NOCS dataset (on the left) and final resulting 3D textured meshes using Hunyuan3D-v2.0 model [28] (on the right).

3.1. Pipeline Architecture

Our pipeline (illustrated in Fig. 1) achieves 3D generation from scratch with minimal human intervention, requiring only 2 hours of manual effort to produce the complete NOCS3D dataset 3. The pipeline has also been validated through 153 categories of the Omni6Dpose [36] dataset as seen in Fig. 2. Our process comprises four phases:

1. LLM-based Geometry Prompt Engineering: Large language models generate category-specific prompts with

Table 2. Comparison of existing category-level 6D pose estimation methods, including quantitative metrics of 3D data usage. *Rast: rasterization; RT: ray tracing; R: real data; MR: Mixed-reality.*

Cat6D dataset	3D dataset	Rend.	#Cat	#O/C	#O	#Img	Code
NOCS-CAMERA25 [29]	ShapeNet	Rast	6	180.8	1080	300K	✗
NOCS-REAL275 [29]	3D scanned	R	6	7	42	8K	N/A
Phocal [31]	3D scanned	R	8	7.5	60	3.9K	N/A
Wild6D [13]	3D scanned	R	5	344.4	1720	10K	N/A
HouseCat6D [18]	3D scanned	R	10	19	190	23.5K	N/A
Omni6DPose-SOPE [36]	OmniObject3D	RT+MR	149	27.9	4023	475K	✗
Omni6DPose-ROPE [36]	3D scanned	R	149	566	3.8	332K	N/A
Omni6D [38]	OmniObject3D	RT	166	28.2	4648	0.8M	✓
Omni6D-xl [38]	Multiple	RT	**419**	38.1	15922	1.1M	✓
Omni6D-real [38]	3D scanned	R	39	1.87	73	1K	N/A
GenNocs6D (Ours)	GenNocs3D	RT+MR	6	**1000**	6000	600K	✓
GenOmni6D (Ours)	GenOmni3D	RT+MR	153	**1000**	**153K**	**1.2M**	✓

Table 3. Pose consistency (%) across methods: depth conditioning impact.

Category	Depth (Ours)	Text-only [10]
Bottle	**100**	70
Bowl	**100**	70
Camera	**97**	30
Can	**100**	70
Laptop	**90**	20
Mug	**100**	82
Avg GenNOCS	**97**	57
Avg GenOmni3D	**96**	–

randomized shape descriptions, while performing self-verification for realism and category coherence.

2. Image Generation and Depth Estimation: The generated prompts produce initial images per category using diffusion models [11]. Manual filtering (<10 minutes per category) removes outliers while preserving geometric diversity and ensuring pose consistency for subsequent 3D mesh generation. DepthFM [14] processes selected images to create depth maps that condition subsequent stages, ensuring pose consistency as shown in Fig. 3.

3. Texture Variation: Each depth map conditions the generation of 10 textured instances through additional LLM texture prompts, yielding 1K total images per category. This approach maintains pose consistency while maximizing realistic shape and texture variations, providing high reliability in text-to-image generation at scale (see Fig. 1 and 3).

4. 3D Reconstruction: A state-of-the-art image-to-3D model is applied to obtain consistently aligned 3D meshes (see Fig. 3), ready for 3D learning, 6D pose estimation, and robotics applications.

3.2. Controlled Image Generation for 3D Pose Consistency

In most image-to-3D pipelines, the orientation of the generated mesh is directly determined by the viewpoint of the input image. As a result, ensuring consistent object poses across image generations is critical for producing canonically aligned 3D meshes. Although some image-to-3D models integrate canonical alignment [28], this typically provides only 90-degree rotational alignment, which is insufficient for global category alignment, thereby necessitating precise image-level pose control.

However, achieving reliable image-level pose consistency remains challenging. Our experiments reveal that text-conditioned generation achieves approximately 80% pose consistency for symmetric objects (e.g., bottles, bowls), but this drops sharply to as low as 20% for complex asymmetric objects such as laptops and cameras. This inconsistency necessitates generating over 5,000 images to obtain just 1,000 usable instances, resulting in substantial computational overhead.

We identify two key limitations: (1) current diffusion models lack explicit understanding of 3D structure during text-to-image generation, and (2) natural language prompts provide ambiguous pose specifications. While prompt engineering can occasionally improve consistency, it often introduces unrealistic artifacts and fails to resolve underlying ambiguities. To address these challenges, we adopt ControlNet [37] with depth-based conditioning. This approach enforces pose consistency during generation and achieves nearly 100% success for simple symmetric objects, while dramatically improving consistency for complex shapes—reaching success rates up to 96%.

As shown in Table 3, using depth-conditioned images increases the average pose consistency across NOCS categories from 57% to 97%. We additionally validated the method on 153 categories from the Omni6Dpose benchmark [36], achieving 96% mesh pose consistency and demonstrating the category-agnostic nature of our generation approach. We also evaluated alternative conditioning methods, such as Canny edge detection, but often introduced visual artifacts due to missing or extraneous edges while preserving unnecessary texture details. In contrast, depth maps proved to be the most effective conditioning signal: they capture global object structure and spatial positioning without constraining local geometric variations or texture details. This enables the generation of images with both pose consistency and geometric diversity (Fig. 3).

3.3. Comparison of image to 3D mesh generation

Recent advances in single-image 3D mesh generation have enabled rapid creation of 3D assets for synthetic datasets. For our scalable applications, we prioritize three key criteria: (i) generation speed (under 1 minute per mesh); (ii) output quality suitable for robotic grasping tasks; and (iii) reliability across diverse object categories. We

Table 4. Comparison of state-of-the-art 3D mesh generation methods. We have evaluated four approaches based on output quality (Qlty), generation time (Time), and visual results (Mesh).

Method	FS3D [1]	SPAR3D [16]	InstantMesh [35]	Hunyuan3D-v2.0 [28]
Quality	⋆	⋆	⋆⋆	⋆⋆⋆
Time	<10s	<10s	<1m	<1m
Mesh				

benchmarked four recent methods meeting our time constraints: FS3D [1], SPAR3D [16], InstantMesh [35], and Hunyuan3D-v2.0 [28]. As shown in Table 4, our evaluation reveals distinct trade-offs: FS3D and SPAR3D offer the fastest generation (<10s) but produce inconsistent quality, particularly for concave objects like mugs; InstantMesh provides improved quality at <1min generation time but struggles with non-convex geometries critical for robotic grasping; while Hunyuan3D-v2.0 achieves the best overall quality and strong reliability in large-scale generation, eliminating the need for manual mesh filtering. It is worth noting that recent 3D mesh generation methods, including Hunyuan3D-v2.0, integrate canonical alignment capabilities. However, these typically provide only 90° rotational alignment, which is insufficient for precise category-based alignment, making our depth-conditioning approach particularly relevant. Furthermore, depth conditioning significantly improves the reliability of textured image generation, nearly eliminating diffusion model failures that commonly occur in text-only methods.

4. Mesh integration for category-level 6D pose dataset generation

This section describes our framework for integrating generated 3D meshes into BlenderProc [8], a widely used Blender-based tool for generating synthetic training data in category 6D pose estimation [15]. Our implementation extends the simulation pipeline from Omni6D [38] (Section 4.1) while also open-source the mixed-reality rendering approach introduced in Omni6DPose [36](Section 4.2). The two methods enable the generation of a category-based 6D pose dataset in a common format with all the needed annotations: RGB-D, instance and semantic masks, NOCS maps[29], 6D poses.

4.1. Complete 3D simulation approach

The first approach, adopted by Omni6D [38], uses full, realistic synthetic scenes from homes scanned in the real world. In this setup, the full-texture 3D scanned scene is loaded into the simulation. Next, we randomly placed objects in physically delineated areas and sampled 10 different camera viewpoints in relation to these object configurations. All scenes are illuminated by five random light sources with random lighting intensity, ensuring sufficient variation in appearance and lighting conditions. Using these methods, we generated a dataset of 300,000 images corresponding to the original size of the NOCS dataset.

4.2. Mixed reality rendering pipeline

The second approach uses a mixed reality approach based on the framework introduced in NOCS [29] and enhanced by Omni6DPose [36]. While the original NOCS implementation placed synthetic objects on planar surfaces against real image backgrounds without shadows, the improved Omni6DPose version incorporates ray-traced shadows with real scanned objects, greatly enhancing scene realism.

Our implementation begins by modeling the physical scene in BlenderProc [8], aligning the camera positions(see Fig. 8) with the background image's viewpoint [2, 29]. Next, we placed objects depending on available planar surfaces using multi-view camera ray casting. To ensure physically plausible poses, the objects are placed in a natural pose and gravity is applied in simulation. Moreover, to generate realistic composite images with objects and shadows overlaid on the background, we employed a technical solution where the simulated scene was rendered invisible while still capturing shadows cast by the objects. This produces a rendered image containing only the objects and their shadows, which is then combined with the background image and real depth, as illustrated in Fig. 7. Additionally, to address depth sensor limitations in capturing dark objects, we have augmented the depth map by integrating partial synthetic depth data, ensuring a complete, object-coherent, and realistic depth representation. Finally, we used this pipeline to generate and release a dataset of 300,000 images corresponding to the original size of the NOCS dataset and more than 1M for the Omni6D pose dataset and finally 475K for the Omni6Dpose dataset.

5. Grasping dataset generation

This section describes the integration of our generated objects into a physical robotic simulation environment, demonstrating our ability to generate a custom grasping dataset from scratch. We selected SAPIEN [34], an open-source, state-of-the-art robotics simulator that provides realistic depth and image rendering through ray tracing. To evaluate grasping baselines using our generated objects and leverage their category-based alignment, we used CenterGrasp [4]. This method utilizes Signed Distance Functions for Grasping (SDFG) to simultaneously train 6D pose estimation and mesh reconstruction from scene RGB-D observations, while also generating grasp poses. Furthermore, CenterGrasp has demonstrated successful zero-shot transfer to real-world scenarios, achieving end-to-end object detection and grasp prediction from single RGB-D inputs using only synthetic training data. Their results show that inte-

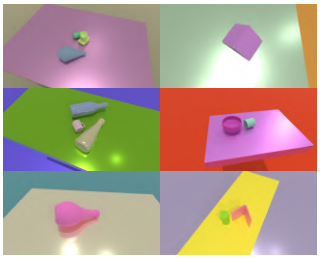

Figure 4. Images with random textures.

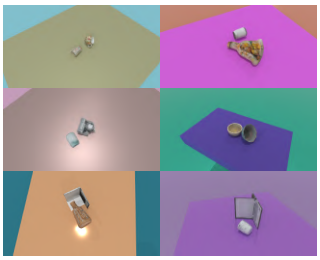

Figure 5. Images with object textures.

Figure 6. Syn images of CAMERA25 [29].

Figure 7. Mixed-reality IKEA dataset (Ours).

Figure 8. REPLICA-based dataset (Ours).

Figure 9. Real images of REAL275 [29].

grating 6D pose estimation improves scene understanding for both reconstruction and grasping tasks. We selected this approach based on the hypothesis that aligned meshes facilitate better 6D pose estimation, making it particularly suitable as a 6D pose-oriented grasping baseline.

5.1. Physical properties

The first step in integrating our generated meshes from Sec 3 into robotic physical simulation software involves assigning realistic physical properties. Similarly to the category-based 6D pose estimation, we established appropriate object scales and sample densities from category-specific ranges to ensure physical plausibility. Additionally, we generated collision meshes from the visual textured meshes to optimize physical collision computations. For this purpose, we used the established V-HACD [21] algorithm for precise convex decomposition. Due to the computational demands of dataset creation, we have limited our dataset to 100 objects per category, resulting in a total of 600 meshes with the associated URDF files.

5.2. SAPIEN scene generation

For scene generation and data preparation, we followed the methodology of CenterGrasp [4], which involves creating two types of scene: ”pile” and ”packed” scenarios. This approach provides diversity in scene complexity, ranging from sparse arrangements with few separated objects to dense, cluttered configurations with stacked objects. For SDFG training data, we generated grasp poses for each mesh. For RGB training, we synthesized images with all necessary annotations, including heatmaps, 6D poses, and latent codes. To facilitate zero-shot transfer to real-world environments, we applied randomizations to ground and table materials and textures across different scenes. Furthermore, to evaluate the impact of our generated textures, compared to the randomized textures (discussed in 6.2), we created two distinct versions of the dataset, as shown in Fig. 4-5.

6. Experiments

This section presents our experimental evaluation of our generated meshes and the dataset generation pipeline on two key tasks: 6D pose estimation (Cat6DPose) and robotic grasping, with particular focus on Sim2Real applications.

6.1. Benchmarking 6D pose Generation on NOCS

This section presents a systematic evaluation of our data generation pipeline by benchmarking its Sim2Real transfer performance on the NOCS REAL275 test set. We adopt DualPoseNet [19] as our baseline, given its strong performance on related benchmarks like Omni6D [38]. To dissect the contribution of different components, we generate and compare five distinct dataset variants, each containing approximately 100K images, to address three key questions:

(i) Mixed-Reality vs. Fully Synthetic for Sim2Real: Our mixed-reality setup (Sec. 4.2) demonstrates a clear advantage over purely synthetic environments (Sec. 4.1). As shown in the top section of Table 10, training with mixed-reality data (Mix^{sh}_{SAI}) achieves the best zero-shot Sim2Real performance on REAL275, with an average score of **34.75**, outperforming the fully synthetic Replica_train baseline (33.10). This validates that our mixed-reality approach provides a more effective bridge to the real world.

(ii) Impact of Generated Mesh Quality: Replacing the

Table 5. Comprehensive evaluation of DualPoseNet for Sim2Real transfer and in-domain performance. The top section reports **zero-shot transfer** to the real-world NOCS REAL275 test set. The middle section shows performance on **synthetic validation** splits. The bottom section provides the original NOCS **supervised upper-bounds** for reference. Metrics evaluate both 2D detection (IoU$50/75$) and 3D pose accuracy, where n°, mcm measures the percentage of poses with rotation error $< n^\circ$ *and* translation error $< m$cm. Best scores for the zero-shot Sim2Real and our synthetic validation experiments are underlined; overall best (supervised) results are in **bold**. The metrics are symmetry aware as done in Omni6D [38]

Training	**Test**	IoU_{50}	IoU_{75}	$5°2\,cm$	$5°5\,cm$	$10°2\,cm$	$10°5\,cm$	$5°$	$10°$	$2\,cm$	$5\,cm$	Avg
Replica_train	Real275 [30]	82.03	34.33	3.40	4.94	11.51	17.12	5.75	18.66	**55.24**	**98.05**	33.10
Mix^{sh}_{SAI}	Real275 [30]	**85.91**	**35.42**	4.62	**7.36**	**13.95**	**22.19**	**8.49**	**23.83**	49.29	96.43	**34.75**
Mix^{no-sh}_{SAI}	Real275 [30]	84.89	28.86	2.27	3.77	8.07	14.13	4.49	15.70	48.43	97.66	30.83
Mix^{sh}_{syn}	Real275 [30]	82.68	31.50	**4.80**	6.32	12.90	17.81	7.01	19.15	52.46	97.77	33.24
Mix^{no-sh}_{syn}	Real275 [30]	73.56	26.05	2.23	3.46	6.81	11.02	4.08	12.31	53.16	**98.05**	29.07
Replica-train	Replica-val	66.43	16.56	1.52	3.04	4.18	8.57	3.34	9.41	**28.05**	**83.67**	22.48
Mix^{sh}_{SAI}_train	Mix^{sh}_{SAI}_val	**69.60**	**22.44**	**2.09**	**4.25**	**5.25**	**10.77**	**4.78**	**12.12**	26.31	81.47	**23.91**
Mix^{no-sh}_{SAI}_train	Mix^{no-sh}_{SAI}_val	68.90	21.72	2.18	4.15	5.16	10.28	4.59	11.41	26.30	81.28	23.60
Mix^{sh}_{syn}_train	Mix^{sh}_{syn}_val	46.93	6.32	0.98	1.74	1.81	3.81	1.83	4.16	16.39	72.59	15.66
Mix^{no-sh}_{syn}_train	Mix^{no-sh}_{syn}_val	46.47	6.46	0.85	1.68	1.70	3.90	1.74	4.22	17.63	72.45	15.71
REAL275_train [29]	REAL275 [29]	79.43	36	32.24	39.24	52.23	67.70	41.44	70.14	70.47	96.89	55.42
CAMERA25 [29] + REAL275 [29]_train	REAL275 [29]	**84.19**	**76**	**40.49**	**45.87**	**63.77**	**75.06**	**47.93**	**76.27**	**82.76**	**99.34**	**61.67**
GenNOCS_train (ours)	REAL275 [29]	81.25	58	35.67	42.15	58.92	71.38	44.28	73.45	76.84	97.89	58.49
GenNOCS_train (ours)	GenNOCS_val	**95.51**	**82.46**	**56.11**	**57.31**	**67.81**	**69.91**	**60.14**	**71.66**	**93.73**	**99.04**	**77.15**
GenOmni3D_train (ours)	GenOmni3D_val	**81.83**	**37.80**	**11.99**	**15.36**	**23.98**	**32.06**	**16.92**	**35.75**	**52.09**	**84.23**	**38.58**

original NOCS synthetic objects with our high-quality generated meshes yields significant improvements. This is evident in both synthetic validation and Sim2Real settings. On synthetic validation, the average metric for our meshes (Mix^{sh}_{SAI}_val) rises to **23.91**, a substantial gain over the **15.66** achieved with original synthetic meshes (Mix^{sh}_{syn}_val). For Sim2Real, our meshes also lead to better performance, e.g., Mix^{no-sh}_{SAI} (30.83) vs. Mix^{no-sh}_{syn} (29.07). We attribute this to the higher fidelity of our assets and the avoidance of categorical inconsistencies present in the original dataset.

(iii) The Role of Shadows in Sim2Real: The presence of shadows in training data is critically important. For both NOCS objects and our meshes, the shadow-enabled versions ($...^{sh}$) consistently and significantly outperform their shadow-free ($...^{no-sh}$) counterparts. For instance, adding shadows improves the average Sim2Real score from 29.07 to 33.24 for NOCS meshes and from 30.83 to 34.75 for our meshes. This underscores that modeling real-world lighting phenomena, like shadows, is essential for closing the Sim2Real gap.

Finally, when trained on our full synthetic dataset and using standard refinement methods, our models achieve zero-shot transfer performance comparable to models trained on real data or combinations of real and synthetic data—remarkably, without using any real training data ourselves. The validation results, provided for benchmarking purposes, confirm the effectiveness of our approach. In summary, the results in Table 10 validate our generation pipeline, demonstrating that high-quality assets rendered in a mixed-reality context with realistic lighting constitute a superior data source for training robust category-level 6D pose estimators.

6.2. Grasping and shape completion evaluation

Table 6. Comparison of our method (*Custom-CG*) with *CenterGrasp* [4] and *GIGA* [17] on grasping and shape completion. Metrics: **Grasp** = grasp success rate (↑), **bi** = bidirectional surface error (↓), **IoU** = Intersection over Union of voxelized meshes (↑). Texture configurations: `tex` = native, `rdom` = randomized, `val` = validation set. E.g., `Custom-CG-tex-val-rdom` is trained with native textures and evaluated with randomized ones. Top four rows: evaluation with textured objects; bottom four: evaluation with randomized textures.

Method	**Grasp success**	**Shape compl.**	
		bi ↓	**IoU** ↑
GIGA-val-tex [17]	0.6375	55.2	0.146
Centergrasp-val-tex [4]	0.7896	27.0	0.314
Custom-CG-rdom-val-tex (Ours)	0.8370	25.9	0.405
Custom-CG-tex-val-tex (Ours)	**0.8679**	**23.5**	**0.475**
GIGA-val-rdom [17]	0.6164	63.9	0.121
Centergrasp-val-rdom [4]	0.8271	28.0	0.312
Custom-CG-tex-val-rdom (Ours)	0.8264	22.3	0.425
Custom-CG-rdom-val-rdom (Ours)	**0.8784**	**19.2**	**0.453**

We evaluated our approach within the CenterGrasp framework [4], using the SAPIEN simulator [34] to assess both shape reconstruction quality and robotic grasping performance with our generated 3D objects. To validate the effectiveness of our textured meshes for custom manipulation tasks, we compare two models trained on our data: one using randomized object textures and another using the native textures generated by our pipeline. Each model is tested in two synthetic environments: (1) with randomized textures and (2) with original textures, allowing a controlled analy-

sis of the impact of the texture on grasping success.

To further demonstrate the advantages of our specific dataset generation method, we benchmark our models against two baselines: (i) the original pre-trained CenterGrasp model [4], trained on over 300 manually curated objects, and (ii) the GIGA model [17], which directly infers grasps from point cloud representations and was also trained on a similarly sized object set. This comparative study highlights the benefits of our tailored, category-aligned dataset in improving both grasp prediction and reconstruction performance.

As shown in Table 9, our custom-generated datasets lead to significant improvements in robotic grasping performance compared to existing baselines in specific tasks. Models trained on our data (Custom-CG variants) outperform both the pre-trained CenterGrasp [4] and GIGA [17] models across all evaluated metrics. In particular, our best-performing model achieves a grasp success rate of 87.8%, surpassing CenterGrasp (82.7%) and GIGA (63.8%). Most notably, shape completion accuracy, measured via bidirectional point-wise error (bi) and Intersection over Union (IoU), improves substantially with our meshes—achieving up to 0.475 IoU, well above the CenterGrasp (0.314) and GIGA (0.146) baselines. These results demonstrate that training on our canonically aligned and textured meshes enables superior generalization and physical realism in both grasping and reconstruction tasks.

6.3. Real robot application

Table 7. Detection and Grasping Success Rate Across Categories

Method	Can	Bottle	Bowl	Mug	Camera	Laptop
Detec (Baseline)	83%	50%	28%	58%	73%	0%
Detec (NOCS)	**100%**	**71%**	71%	**83%**	60%	**100%**
Detec (Specialized)	**100%**	66%	**86%**	75%	**100%**	**100%**
Grasp (Baseline)	96%	50%	14%	50%	66%	0%
Grasp (NOCS)	**100%**	46%	47%	58%	60%	0%
Grasp (Specialized)	**100%**	**52%**	**57%**	**71%**	**73%**	0%

We conduct a zero-shot sim-to-real benchmark of the CenterGrasp model on objects from the six NOCS categories: Can, Bottle, Bowl, Mug, Camera, and Laptop. This evaluation assesses detection and grasping performance across three model variants: the original Baseline model pre-trained on 300 diverse objects; our general-purpose NOCS model, trained on a custom dataset aligned with the NOCS categories; and our Specialized models, which are fine-tuned on category-specific data. The evaluation protocol involved 36 unique objects. For each object category, three distinct RGB-D views were captured, resulting in a total of 108 pose estimations. All three models were evaluated on the same input data in a zero-shot manner, ensuring a fair comparison of their object detection and subsequent grasp success rates. The results, summarized in Table 7 and visualized in Fig. 10, demonstrate a clear performance hierarchy. Our NOCS model, trained on custom data, achieves a reliable grasping success rate of over 50%, significantly surpassing the baseline. This confirms the advantage of using tailored datasets for specific object domains. Furthermore, the category-specialized models achieve the highest performance, substantially outperforming both the baseline and the general NOCS model. This underscores the potential of easily generated, highly-specific training data to push the boundaries of real-world robotic grasping performance beyond the capabilities of general-purpose datasets.

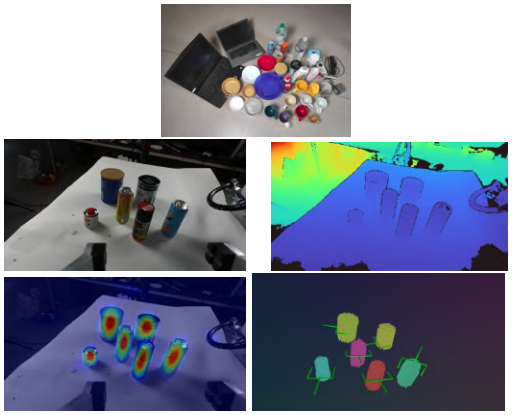

Figure 10. Robotic perception and grasping visualizations on can object of Centergrasp trained on our custom "Can" dataset. Top: All 6 category objects used in real-world setup. Bottom (2×2 grid): top row—RGB image of a scene of cans and associated depth; bottom row—shape-based grasp prediction and heatmap indicating grasp confidence.

7. Conclusion

This work presents an attempt to overcome the critical data bottleneck in 3D vision through an automated pipeline that generates complete category-level 6D pose and grasping datasets directly from text categories. Our key innovation lies in generating high-quality, canonically aligned 3D meshes with 96% pose consistency across 153 diverse categories, while achieving a 5-20× speedup over traditional scanning methods with generation times. The released datasets—comprising the largest aligned 3D mesh collection (153K meshes) and category-level 6D pose dataset (1.2M images)—provide scale for developing 3D foundation models. Extensive validation demonstrates competitive zero-shot sim2real transfer on the NOCS benchmark and superior robotic grasping performance with 87.8% success rate in simulation, confirmed through zero-shot real-world testing. By enabling efficient, scalable generation of task-specific 3D data from scratch, our work transforms the paradigm of 3D dataset creation and paves the way for accelerated progress in 3D vision and robotic manipulation.

8. Acknowledgment

We sincerely thank Mengchen Zhang (author of Omni6D [38] for her valuable assistance in using BlenderProc and reproducing baseline results on our custom dataset. We also extend our gratitude to Jiyao Zhang (co-author of [36]) for his guidance in implementing the Mixed Reality pipeline. We are grateful to Tencent for providing access to Hunyuan3D 2.0, enabling the generation and open sharing of our dataset for our research applications.

This work was in part supported by the French Research Agency, l'Agence Nationale de Recherche (ANR), through the projects Learn Real (ANR-18-CHR3-0002-01), Chiron (ANR-20-IADJ-0001-01), Aristotle (ANR-21-FAI1-0009-01), Astérix (ANR-23-EDIA-0002), Demeter (ANR-25-HTCE-0002) and Protheus (ANR-25), the French national investment prioritary program PSPC FAIR WASTE project, as well as a donation to Fonds de Dotation Centrale Lyon by Huawei Technologies R&D France. It was granted access to the HPC resources of IDRIS under the allocation 2025-[AD011015271R1], 2025-[AD011015591R1] and 2026-[A0191013894] made by GENCI.

References

[1] Mark Boss, Zixuan Huang, Aaryaman Vasishta, and Varun Jampani. Sf3d: Stable fast 3d mesh reconstruction with uv-unwrapping and illumination disentanglement. *arXiv preprint*, 2024. 5

[2] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from rgb-d data in indoor environments. *arXiv preprint arXiv:1709.06158*, 2017. 5

[3] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015. 2, 3

[4] Eugenio Chisari, Nick Heppert, Tim Welschehold, Wolfram Burgard, and Abhinav Valada. Centergrasp: Object-aware implicit representation learning for simultaneous shape reconstruction and 6-dof grasp estimation. *IEEE Robotics and Automation Letters*, 2024. 5, 6, 7, 8

[5] Jasmine Collins, Shubham Goel, Kenan Deng, Achleshwar Luthra, Leon Xu, Erhan Gundogdu, Xi Zhang, Tomas F Yago Vicente, Thomas Dideriksen, Himanshu Arora, Matthieu Guillaumin, and Jitendra Malik. Abo: Dataset and benchmarks for real-world 3d object understanding. *CVPR*, 2022. 3

[6] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13142–13153, 2023. 3

[7] 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. *Advances in Neural Information Processing Systems*, 36, 2024. 2, 3

[8] Maximilian Denninger, Dominik Winkelbauer, Martin Sundermeyer, Wout Boerdijk, Markus Knauer, Klaus H. Strobl, Matthias Humt, and Rudolph Triebel. Blenderproc2: A procedural pipeline for photorealistic rendering. *Journal of Open Source Software*, 8(82):4901, 2023. 5

[9] 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 *2022 International Conference on Robotics and Automation (ICRA)*, pages 2553–2560. IEEE, 2022. 3

[10] Guillaume Duret, Younes Bourenname, Danylo Mazurak, Anna Samsonenko, Florence Zara, Liming Chen, and Jan Peters. Facilitate and scale up the creation of 3d meshes and 6d category-based datasets with generative models: Genvegefruits3d. *HAL*, 2025. 2, 3, 4

[11] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024. 4

[12] Huan Fu, Rongfei Jia, Lin Gao, Mingming Gong, Binqiang Zhao, Steve Maybank, and Dacheng Tao. 3d-future: 3d furniture shape with texture. *International Journal of Computer Vision*, pages 1–25, 2021. 3

[13] Yang Fu and Xiaolong Wang. Category-level 6d object pose estimation in the wild: A semi-supervised learning approach and a new dataset. *Advances in Neural Information Processing Systems*, 35: 27469–27483, 2022. 3, 4

[14] Ming Gui, Johannes Schusterbauer, Ulrich Prestel, Pingchuan Ma, Dmytro Kotovenko, Olga Grebenkova, Stefan Andreas Baumann, Vincent Tao Hu, and Björn Ommer. Depthfm: Fast generative monocular depth estimation with flow matching. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3203–3211, 2025. 4

[15] Tomas Hodan, Martin Sundermeyer, Yann Labbe, Van Nguyen Nguyen, Gu Wang, Eric Brachmann, Bertram Drost, Vincent Lepetit, Carsten Rother, and Jiri Matas. Bop challenge 2023 on detection segmentation and pose estimation of seen and unseen rigid objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5610–5619, 2024. 5

[16] Zixuan Huang, Mark Boss, Aaryaman Vasishta, James M Rehg, and Varun Jampani. Spar3d: Stable point-aware reconstruction of 3d objects from single images. *arXiv preprint arXiv:2501.04689*, 2025. 5

[17] Zhenyu Jiang, Yifeng Zhu, Maxwell Svetlik, Kuan Fang, and Yuke Zhu. Synergies between affordance and geometry: 6-dof grasp detection via implicit representations. *arXiv preprint arXiv:2104.01542*, 2021. 7, 8

[18] HyunJun Jung, Shun-Cheng Wu, Patrick Ruhkamp, Guangyao Zhai, Hannah Schieber, Giulia Rizzoli, Pengyuan Wang, Hongcheng Zhao, Lorenzo Garattoni, Sven Meier, et al. Housecat6d-a large-scale multi-modal category level 6d object perception dataset with household objects in realistic scenarios. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22498–22508, 2024. 3, 4

[19] Jiehong Lin, Zewei Wei, Zhihao Li, Songcen Xu, Kui Jia, and Yuanqing Li. Dualposenet: Category-level 6d object pose and size estimation using dual pose network with refined learning of pose consistency. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3560–3569, 2021. 6

[20] Liu Liu, Wenqiang Xu, Haoyuan Fu, Sucheng Qian, Qiaojun Yu, Yang Han, and Cewu Lu. Akb-48: A real-world articulated object knowledge base. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14809–14818, 2022. 3

[21] Khaled Mamou and Faouzi Ghorbel. A simple and efficient approach for 3d mesh approximate convex decomposition. In *2009 16th IEEE international conference on image processing (ICIP)*, pages 3501–3504. IEEE, 2009. 6

[22] Adithyavairavan Murali, Balakumar Sundaralingam, Yu-Wei Chao, Jun Yamada, Wentao Yuan, Mark Carlson, Fabio Ramos, Stan Birchfield, Dieter Fox, and Clemens Eppner. Graspgen: A diffusion-based

framework for 6-dof grasping with on-generator training. *arXiv preprint arXiv:2507.13097*, 2025. 2

[23] Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2023. 1

[24] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[25] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 1

[26] Stefan Stojanov, Anh Thai, and James M. Rehg. Using shape to categorize: Low-shot learning with an explicit shape bias. *CVPR*, 2021. 3

[27] Julian Straub, Thomas Whelan, Lingni Ma, Yufan Chen, Erik Wijmans, Simon Green, Jakob J. Engel, Raul Mur-Artal, Carl Ren, Shobhit Verma, Anton Clarkson, Mingfei Yan, Brian Budge, Yajie Yan, Xiaqing Pan, June Yon, Yuyang Zou, Kimberly Leon, Nigel Carter, Jesus Briales, Tyler Gillingham, Elias Mueggler, Luis Pesqueira, Manolis Savva, Dhruv Batra, Hauke M. Strasdat, Renzo De Nardi, Michael Goesele, Steven Lovegrove, and Richard Newcombe. The Replica dataset: A digital replica of indoor spaces. *arXiv preprint arXiv:1906.05797*, 2019. 3

[28] Tencent Hunyuan3D Team. Hunyuan3d 2.0: Scaling diffusion models for high resolution textured 3d assets generation, 2025. 3, 4, 5

[29] He Wang, Srinath Sridhar, Jingwei Huang, Julien Valentin, Shuran Song, and Leonidas J. Guibas. Normalized object coordinate space for category-level 6d object pose and size estimation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 3, 4, 5, 6, 7

[30] He Wang, Srinath Sridhar, Jingwei Huang, Julien Valentin, Shuran Song, and Leonidas J. Guibas. Normalized object coordinate space for category-level 6d object pose and size estimation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 7

[31] Pengyuan Wang, HyunJun Jung, Yitong Li, Siyuan Shen, Rahul Parthasarathy Srikanth, Lorenzo Garattoni, Sven Meier, Nassir Navab, and Benjamin Busam. Phocal: A multi-modal dataset for category-level object pose estimation with photometrically challenging objects. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 21222–21231, 2022. 3, 4

[32] 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

[33] Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 3d shapenets: A deep representation for volumetric shapes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1912–1920, 2015. 3

[34] Fanbo Xiang, Yuzhe Qin, Kaichun Mo, Yikuan Xia, Hao Zhu, Fangchen Liu, Minghua Liu, Hanxiao Jiang, Yifu Yuan, He Wang, Li Yi, Angel X. Chang, Leonidas J. Guibas, and Hao Su. SAPIEN: A SimulAted Part-Based Interactive ENvironment. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11094–11104. IEEE, 2020. 2, 5, 7

[35] Jiale Xu, Weihao Cheng, Yiming Gao, Xintao Wang, Shenghua Gao, and Ying Shan. Instantmesh: Efficient 3d mesh generation from a single image with sparse-view large reconstruction models. *arXiv preprint arXiv:2404.07191*, 2024. 5

[36] Jiyao Zhang, Weiyao Huang, Bo Peng, Mingdong Wu, Fei Hu, Zijian Chen, Bo Zhao, and Hao Dong. Omni6dpose: a benchmark and model for universal 6d object pose estimation and tracking. In *European Conference on Computer Vision*, pages 199–216. Springer, 2025. 2, 3, 4, 5, 9

[37] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023. 4

[38] Mengchen Zhang, Tong Wu, Tai Wang, Tengfei Wang, Ziwei Liu, and Dahua Lin. Omni6d: Large-vocabulary 3d object dataset for category-level 6d object pose estimation. In *European Conference on Computer Vision*, pages 216–232. Springer, 2025. 2, 3, 4, 5, 6, 7, 9

Breaking the 3D Dataset Bottleneck: Fast Scalable Generation of Aligned 3D Assets from Scratch for Category 6D Pose Estimation and Robotic Grasping

Supplementary Material

9. Prompt engineering and reference pose choice

In the first step of generating geometry-based images, we position the object in a front-facing view to maximize consistency in the initial image generation. This choice was made because text-to-image models struggle with three-quarter views, often requiring significantly more iterations to produce 100 well-positioned images. However, this approach remains highly object-dependent, with only around 20% of complex objects yielding satisfactory results.

While front-view optimization improves initial image alignment, it impacts the downstream pipeline, particularly in textured image-to-3D conversion. A three-quarter view provides more complete shape information, helping to mitigate common 3D reconstruction issues. This is particularly useful for objects like cameras, where classical diffusion models can still fail to maintain global 3D coherence - for example, by incorrectly generating lenses on both the front and back of the object. Future work could enforce specific viewing angles during image generation to achieve near 100% success in aligned mesh generation.

For geometry and texture generation, we employ LLMs to generate diverse visual descriptions of target categories based on a prompt skeleton. A potential improvement would be integrating LLMs directly into the pipeline, enabling dynamic, randomized descriptions per object rather than relying on static text files. All geometric, visual and validation prompts (shape, texture, color) are generated using a structured template using the `Llama-3.3-70B-Instruct` model. The format follows:`Generate EXACTLY {count} standalone [geometric/visual] descriptions for a {category}.` This instruction is accompanied by strict formatting rules (e.g., `ONLY commaseparated items`) and content guidelines (e.g., `Focus on {category}specific shapes`), each illustrated with clear good and bad examples. The subsequent "selfverification for realism and category coherence" is a separate validation step, performed using an analogous `VISUAL_DESCRIPTION_PROMPT` skeleton. A generated description is rejected if it contradicts typical attributes—for example, `Rejected: Tomatoes are typically red, not soft pink.` Using the final `Llama-3.3-70B-Instruct` model, over 90% of prompts pass this check.

For the 153 categories, given that our input consists of single category labels, we systematically renamed category names to avoid ambiguity in text-to-image generation. Our analysis revealed that 20 out of 153 categories (approximately 13%) required more specific naming to enhance generation quality and reduce errors. The renaming strategy focused on three key improvements:

First, we added specific descriptors to clarify object states and configurations, such as "banana" → "single_banana", "belt" → "rolled_up_belt", and "plug" → "male_electric_plug". These modifications provided the diffusion models with precise contextual information about object orientation and form, enabling more accurate generation.

Second, we enhanced specificity with technical precision, including "sword_bean" → "canavalia_ensiformis_sword_bean" and "boxed_beverage" → "supermarket_boxed_beverage". This technical specificity aligned with the model's training data distribution and reduced ambiguous interpretations, preventing instances where generic prompts like "sword_bean" would incorrectly generate images of metallic weapons rather than the intended legume species.

Third, we incorporated positional and contextual modifiers like "keyboard" → "keyboard_from_upward", and "toy_car" → "toy_car_long_side", which provided crucial spatial relationships that significantly improved generation consistency and reduced perspective errors.

Our empirical validation demonstrated that these semantic enhancements improved generation accuracy by approximately 27% by providing clearer textual guidance to the text encoder. The more descriptive naming convention better leverages the compositional understanding of modern text-to-image models, resulting in more precise object representation and reduced hallucination across all 153 object categories. The increased specificity proved particularly valuable for objects with multiple common configurations or viewing angles, where ambiguous prompts previously led to inconsistent results.

10. Meshes variety discussion

Regarding geometric diversity: first, the depth modality provides a relatively low-resolution, blurry input. Consequently, the generated 3D mesh for a given depth map can vary significantly in its fine-grained geometry due to differences in the random seed and the accompanying textural prompt(mostly illustrated by Figures 12, 11 and the camera object). Second, and more importantly, each generated

mesh is subsequently scaled by a random factor sampled from a category-specific realistic range. This scaling operation produces substantial variation in the absolute size. Regarding potential model failures (e.g., artifacts or pose inconsistencies), we intentionally avoid explicit filtering. Our goal is to demonstrate the pipeline's robustness for large-scale creation. Occasional inconsistent outputs are treated as a form of domain randomization. As the field progresses, any component can be seamlessly upgraded; further refinement, while beneficial, is beyond the scope of this paper.

11. Pose Consistency Evaluation

Pose consistency is evaluated through manual human judgment. This approach is necessary due to the challenges in defining an automated metric for perceptual pose alignment across categories with high shape variation. Annotators visually assessed pairs of reconstructed and reference meshes in a 3D viewer, labeling a pair as consistent only if the reconstruction was both accurate and in the same 3D orientation as the reference. This evaluation was performed on 100 meshes per category from NOCS (600 total) and 10 meshes per category from Omni3D (1530 total).

12. 3D analysis of the 3D meshes

In addition to downstream applications, we conducted a comprehensive qualitative evaluation comparing our generated 3D meshes with real-world scanned meshes from the Omni6DPose dataset. Given the absence of ground-truth 3D geometry for direct comparison, conventional 3D evaluation metrics such as Chamfer distance or FID scores are not directly applicable. Instead, we employ CLIP-based semantic evaluation to assess the perceptual quality and categorical accuracy of the generated assets.

For each object in our OMNI3D dataset and corresponding real-world scans from Omni6DPose, we render six consistent orthographic views (front, back, left, right, top, bottom) under standardized lighting conditions. These multi-view renderings, paired with their respective category labels, are processed using the state-of-the-art CLIP model (ViT-L/32) to compute three key metrics: *Category Score* (classification confidence), *Realism Score* (photorealism), and *Consistency* (multi-view coherence). 1) Category Score uses prompts like ["a photo of a C", "a 3D model of a C", "a C object"], averaging results; 2) Realism Score uses the prompt "a realistic C"; and 3) View Consistency follows [1], using viewpoint-specific prompts ("front/side/back view of a C") to measure alignment across renders. The final score for each metric is the average across all views and prompt variations.

As demonstrated in Table 8, our AI-generated meshes achieve comparable and in some cases superior metrics to real-world scanned meshes. Notably, the AI models exhibit higher *Overall Quality* (0.4493 vs. 0.4490) and *Category Score* (0.2555 vs. 0.2492), indicating better semantic alignment and categorical distinctiveness. The marginal deficit in *Realism Score* (0.1717 vs. 0.1773) is offset by near-perfect *Consistency* across views (0.9206), underscoring the structural coherence of generated assets. These results substantiate that AI-generated 3D meshes can effectively substitute real-world scans while offering the scalability advantages of procedural generation.

Table 8. 3D Mesh Quality Comparison

Metric	Real-World	AI Models
Category	0.2492	**0.2555**
Realism	**0.1773**	0.1717
Consistency	0.9205	**0.9206**
Average	0.4490	**0.4493**

To quantitatively evaluate manipulation suitability, we conducted an experiment taking advantage GraspDataGen[22] parallelized framework in Isaac Sim. We sampled and simulated over 1.5 million parallel-jaw grasps across 10 meshes for each of the 153 Omni3D categories. The overall success rate was 72.4%, demonstrating that generated assets afford stable grasps. The hard failures are explainable and often stem from inherent object geometry (e.g., watermelons) that are unsuitable for a normal-sized parallel gripper, which validates the physical plausibility of the results. We will add these quantitative results (success rates, analysis per category) and release the grasp evaluation.

13. Aligned Mesh impact on Grasping Success

Table 9. Center-grasp results without ICP applied

Scenario	Grasp success ↑	Declutter ↑	Shape compl. bi ↓	Shape compl. IoU ↑
GenBanana (Aligned)	**0.503**	**0.509**	**18.5**	**0.391**
GenBanana (Random)	0.361	0.378	19.2	0.367
GenNOCS (Aligned)	**0.261**	**0.274**	**26.5**	**0.378**
GenNOCS (Random)	0.245	0.257	31.7	0.339

In this section, we evaluate the impact of using aligned meshes for global object-centered pose estimation and shape reconstruction in grasping tasks. Importantly, no Iterative Closest Point (ICP) refinement is applied, so the reported values reflect the raw outputs of the models. This allows us to directly assess how alignment influences learning quality without post-processing corrections.

The evaluation is conducted across three distinct scenarios:

- **Banana Objects:** 100 instances of bananas, chosen for their commonality and challenging non-convex geometry.

Table 10. Performance 6D pose Comparison

Training Size	IoU_{50}	IoU_{75}	$5°2\,cm$	$5°5\,cm$	$10°2\,cm$	$10°5\,cm$	$5°$	$10°$	$2\,cm$	$5\,cm$	Avg
10 objects NOCS3D	93.01	76.48	42.41	45.99	57.63	64.21	51.03	66.71	83.61	97.89	67.80
100 objects NOCS3D	<u>94.00</u>	<u>79.08</u>	<u>47.75</u>	<u>49.97</u>	<u>62.19</u>	<u>66.42</u>	<u>53.52</u>	<u>67.81</u>	<u>89.48</u>	**98.99**	<u>69.92</u>
1000 objects NOCS3D	**94.62**	**81.40**	**48.84**	**50.69**	**64.57**	**67.85**	**54.28**	**69.40**	**91.23**	<u>98.78</u>	**71.07**

- **NOCS Categories:** 100 meshes for each of the six object categories from the NOCS dataset.

Our analysis (Table 9) shows that aligned meshes consistently improve grasp success, decluttering efficiency, and shape reconstruction accuracy compared to randomly oriented inputs. By removing rotational ambiguity, alignment provides a clearer signal for the model to learn object-centered features. This is particularly beneficial for categories with complex geometries (e.g., bananas), where random orientations can obscure structural cues. Overall, alignment enhances the stability of learning and reduces reliance on corrective post-processing, demonstrating its value for robust grasping pipelines.

14. Cat 6D pose at scale

To study the impact of 3D asset scale, we designed a controlled experiment evaluating pose estimation performance by varying the number of unique 3D assets used during training, while keeping the total number of generated synthetic images constant. Table 10 shows that a variety of meshes is beneficial for 6D pose. Beyond immediate pose estimation, we believe the ability to generate large-scale, category-aligned 3D datasets automatically is a key contribution for other models like 3D understanding tasks (e.g., implicit representation learning, 6D pose estimation with 3D priors). To provide quantitative validation of 153cat, we train DualPoseNet on our 153-category dataset and evaluate its performance on the validation set. The results show strong 6D pose estimation performance, with metrics approximately twice as high as those reported by Omni6D. Although limited to validation set evaluation, these results demonstrate that our large-scale generated meshes are suitable for category-level 6D pose.

15. More dataset visualization

This section shows more image visualizations of the generated data in the paper. Starting from Images to 3D meshes, more images of the IKEA [29], REPLICA [27], datasets, and the two centergrasp datasets.

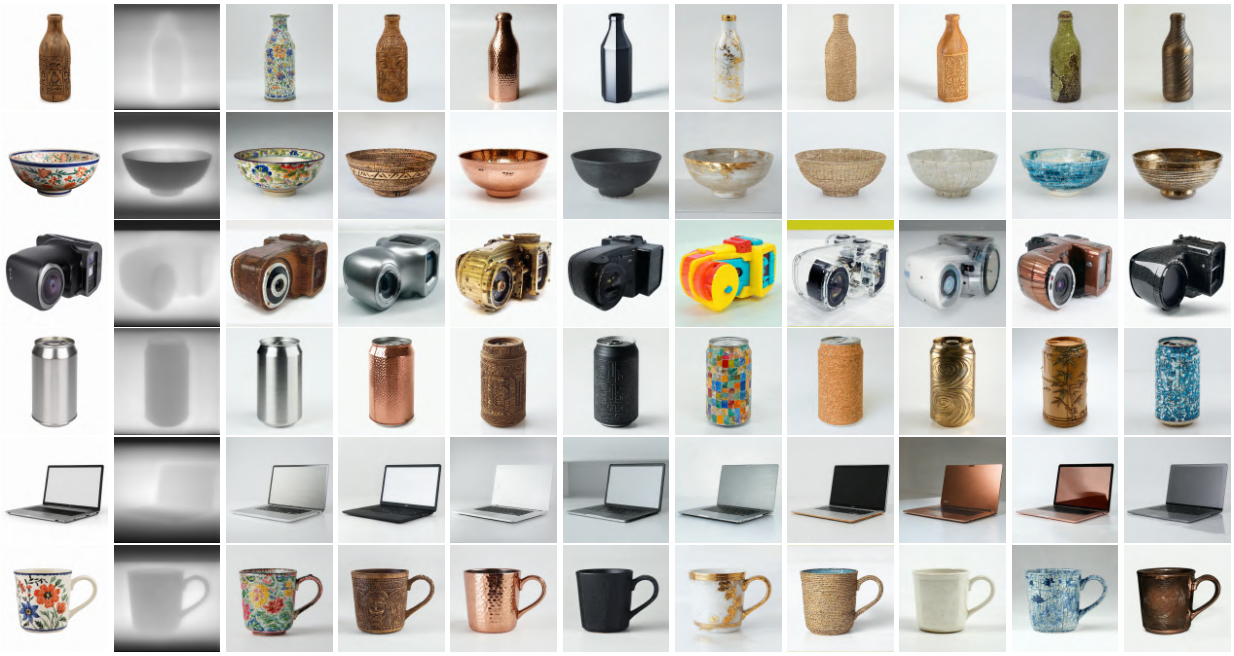

Figure 11. Example of images generated through depth conditioning, the first image is initial image, the second is the depth, the following images are conditioned based textured images with different texture prompts.

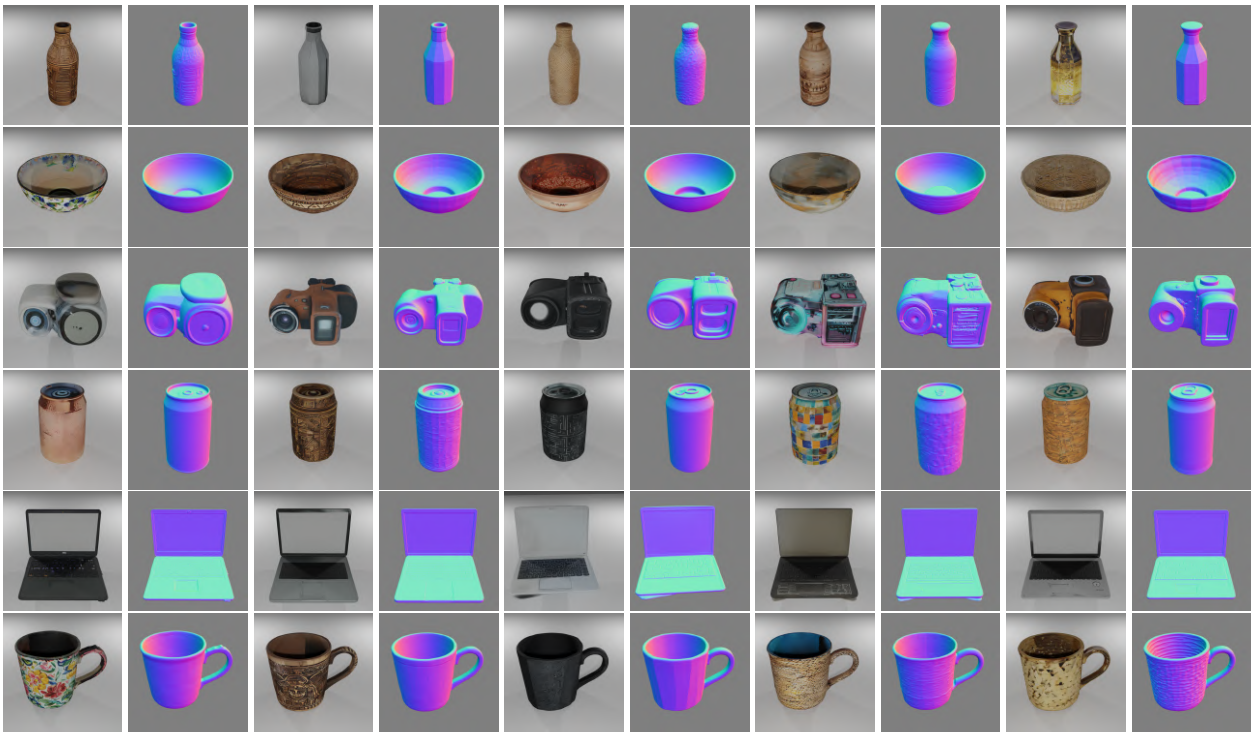

Figure 12. Different input features generated from prompt-based generated images.

Figure 13. NOCS Replica dataset samples showing various household objects and food items. The images demonstrate the diversity of objects in the dataset with consistent lighting and background conditions.

Figure 14. NOCS IKEA dataset samples featuring various IKEA products in different configurations and viewpoints. The images show the variety of household items with different textures and shapes.

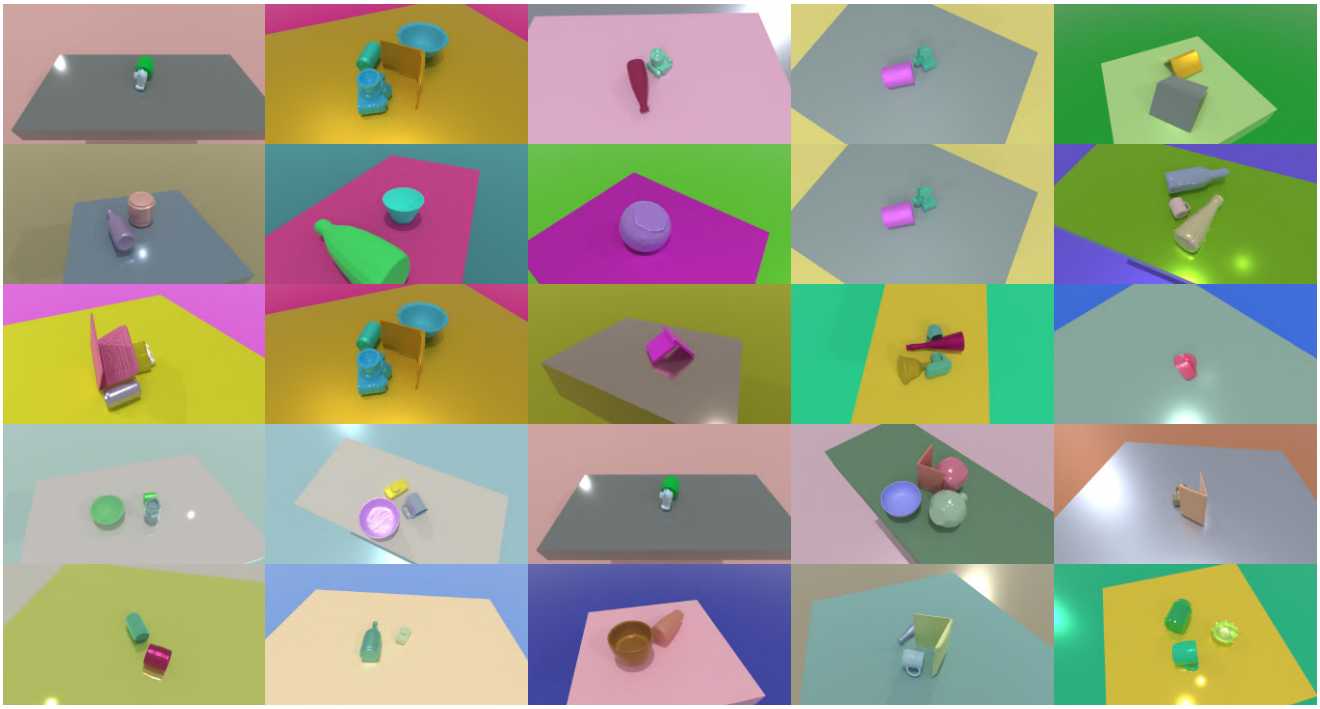

Figure 15. Centergrasp dataset examples of GenNOCS3D object with random textures

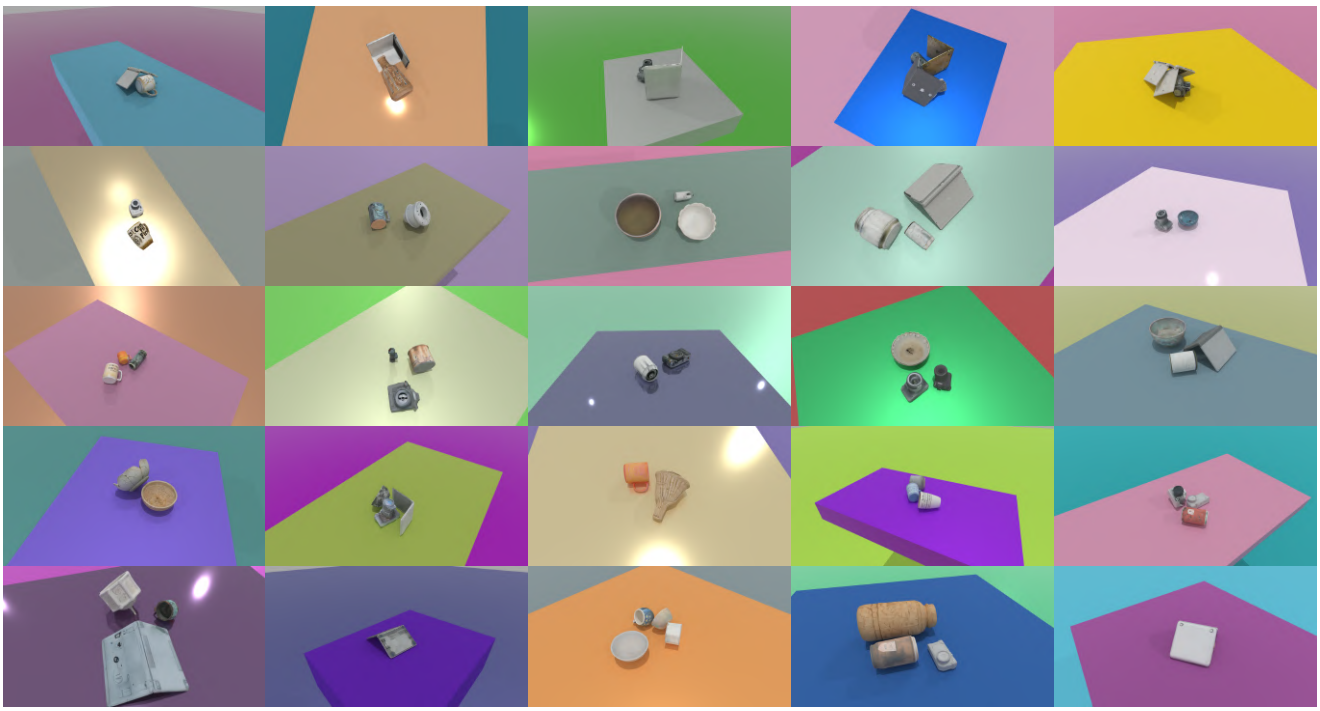

Figure 16. Centergrasp dataset examples of GenNOCS3D object with native textures.

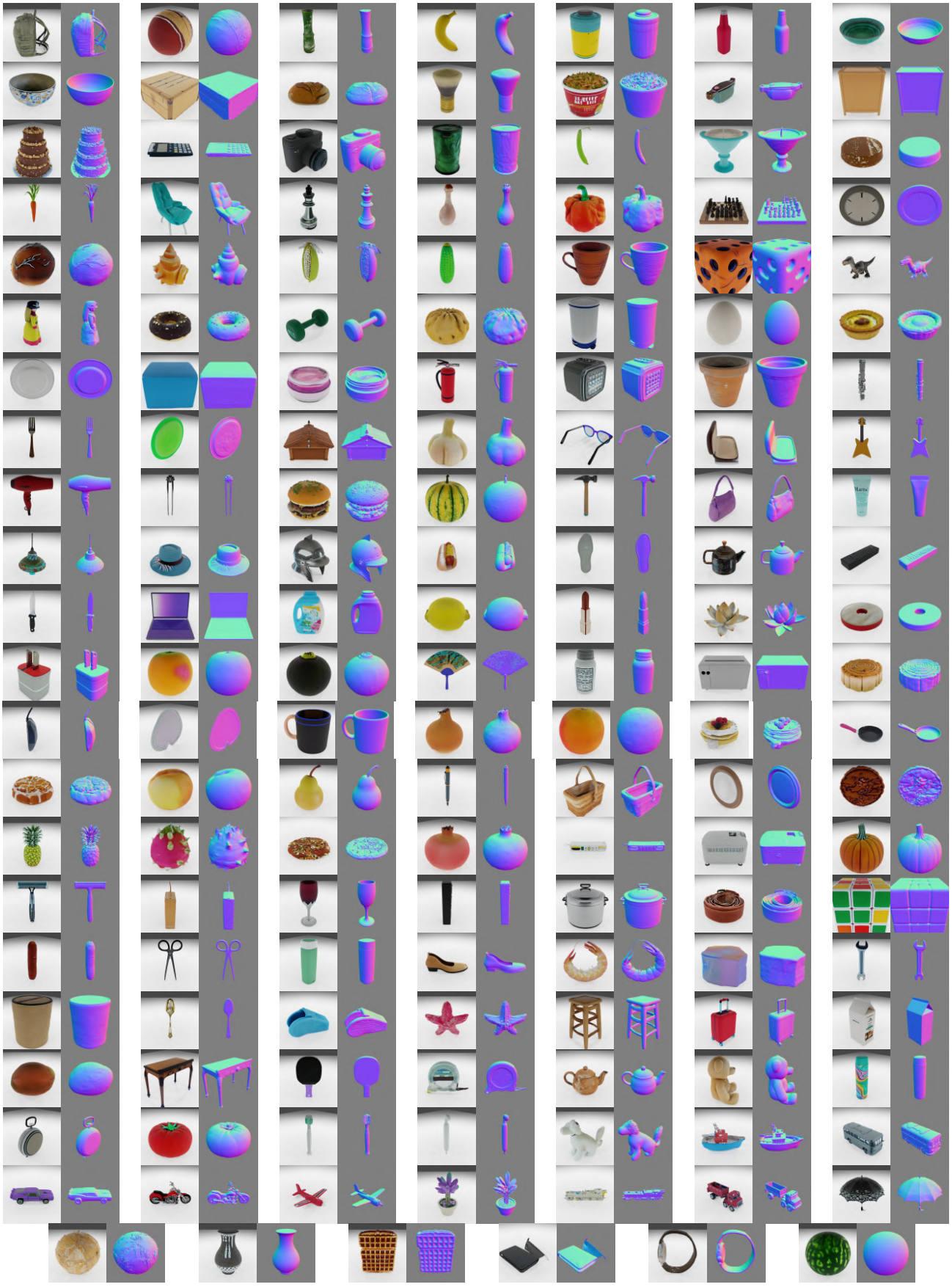

Figure 17. Omni3D Dataset: Large RGB and Normal images for various objects. Each pair shows the RGB image (left) and corresponding normal map (right).