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

Guillaume Duret, Danylo Mazurak, Florence Zara, Jan Peters, Liming Chen

Breaking the 3D Dataset Bottleneck: Fast Scalable Generation of Aligned 3D Assets from Scratch for 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

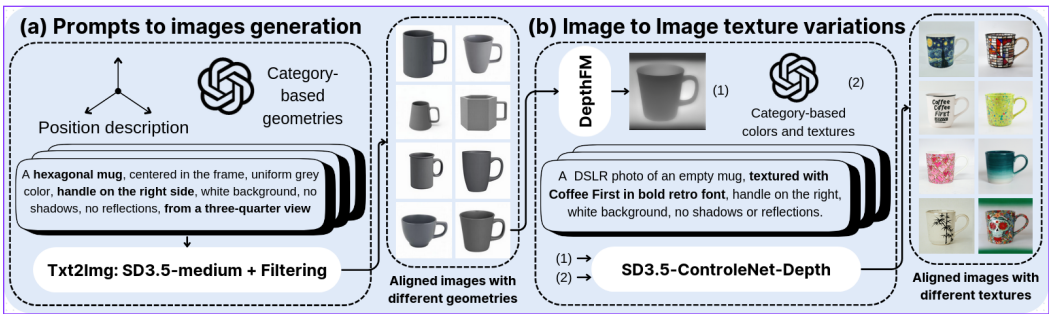


Figure 1: Our text-to-image pipeline: (a) Category-based geometry prompt engineering and images generation; (b) Depth-conditioned image generation for texture variation and automatic alignment.

Abstract

We introduce a scalable framework for generating category-level 6D pose datasets from text prompts automatically. Our approach addresses three key bottlenecks in the creation of 3D datasets: (1) automated asset generation through a controlled text-to-image-to-3D pipeline; (2) built-in canonical alignment through depth-conditioned generation; and (3) large-scale 6D annotation using mixed reality rendering. GenNOCS produces high-quality aligned 3D meshes in under 3min per object, achieving a 5–20× speedup over traditional scanning—with a 96 mesh generation success rate including consistent pose alignment. We demonstrate the applicability of our pipeline by integrating these meshes in state-of-the-art sim2real 6D pose generation. We demonstrate sim2real transfer on the NOCS benchmark for 6D pose estimation with competitive results in a zero-shot manner. Finally, we confirm the practical utility of our generated assets in real-world robotic grasping scenarios. By eliminating dependencies on existing 3D assets, our method enables rapid creation of custom 6D datasets, achieving 6000 aligned instances over 6 categories of the NOCS benchmark with only 20min of human effort. This work provides a critical step toward foundation models for 3D understanding, offering new possibilities for research in 6D perception and manipulation for custom datasets. The code and dataset from prompt to 6D pose dataset generation are publicly available at `https://huggingface.co/datasets/Guillaume0477/GenNOCS`.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

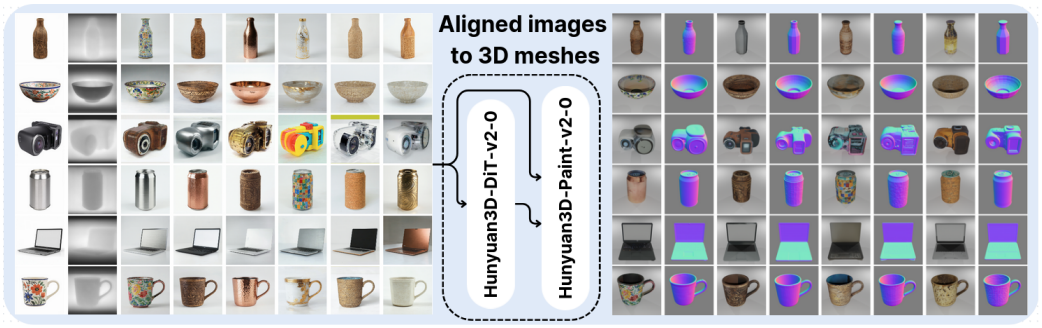


Figure 2: Qualitative examples of final textured images used of the 6 categories of NOCS dataset (on the left) and final resulted 3D textured meshes using Hunyuan3D-v2.0 model [24] (on the right).

1 Introduction

The field of 2D computer vision has seen significant advancements driven by foundation models trained on large-scale datasets. In contrast, progress in 3D vision remains constrained by a critical limitation: the scarcity of high-quality, diverse, and scalable data. Despite this, an advanced 3D understanding is essential for numerous applications, including robotics, augmented reality, and scene reconstruction. Central to this challenge is 6D pose estimation—the task of predicting an object's 3D position and orientation from a single RGB-D image, even in the absence of exact target 3D models. In the context of category-level 6D pose estimation, the reliance on precise object meshes is mitigated. However, this task still requires extensive category-specific datasets comprising aligned 3D assets with significant intra-class shape and texture variation. The creation of such datasets involves three primary challenges:

1. Asset Collection: Existing methods depend on labor-intensive 3D scanning (15–60 minutes per object) or pre-existing 3D repositories. Current 3D datasets can be categorized as follows: (1) synthetic collections that lack realism; (2) high-quality scanned datasets that are limited in scale; and (3) large-scale, internet-sourced repositories with inconsistent mesh quality, poor alignment, and incomplete category labeling. The data scarcity is even more pronounced for category-level 6D pose estimation, where existing datasets cover only a limited number of categories with insufficient instance diversity.

2. Mesh Alignment: Canonical alignment across object instances is crucial for effective training. However, achieving consistent alignment requires extensive manual effort, hindering large-scale dataset creation. While existing datasets provide some aligned assets, the demand for large-scale, consistently aligned datasets remains largely unmet in the 6D pose estimation domain.

3. Pose Annotation: Real-world 6D pose annotation often involves pre-scanned assets and iterative optimization (e.g., ICP), making it error-prone and challenging to scale across categories and image samples. Synthetic datasets offer partial solutions but are limited in their ability to generate diverse scenes, restricting scalability.

To address these challenges, we propose a scalable pipeline for generating realistic, aligned 3D meshes and 6D pose datasets for any object category. Our method facilitates rapid prototyping, robust sim-to-real transfer, and robotic grasping—all without relying on pre-existing 3D models. Our contributions include:

- *Automated Text-to-3D Generation:* We introduce a pipeline that synthesizes aligned, category-specific 3D meshes from text prompts. By leveraging depth-conditioned diffusion models, we achieve over 97% automatic pose alignment with minimal human intervention.
- *Hybrid Rendering for Sim-to-Real Transfer:* To bridge the sim-to-real gap, we incorporate two complementary rendering strategies: (a) mixed-reality composition with physically plausible shadows and (b) full 3D simulation using randomized lighting and viewpoints.
- *Sim-to-Real Benchmarking:* We evaluate the generated datasets on the NOCS benchmark, demonstrating state-of-the-art zero-shot sim-to-real performance in 6D pose estimation. Our ablation studies further validate the superiority of our synthetic data over existing synthetic assets.
- *Robotic Grasping Validation:* We assess the applicability of our generated meshes in the SAPIEN simulator for task-specific dataset generation, showing that our pipeline outperforms models trained on extensive datasets in specialized grasping tasks.

By decoupling dataset creation from manual scanning, our work establishes a foundation for rapid 3D asset generation, facilitating scalable dataset creation for training foundation models for 3D perception and manipulation. We release our code, pipeline, and datasets to support advancements in 3D vision and robotics: `https://huggingface.co/datasets/Guillaume0477/GenNOCS`.

2 Related Work

2.1 3D Datasets

As shown in Table 1, 3D datasets fall into three categories:

Synthetic datasets: ShapeNet [3] (55K objects) established early benchmarks, while later datasets (ModelNet [30], 3D-FUTURE [12], ABO [5], Toy4K [22]) improved variety but lacked photorealism.

Real-world scans: GSO (1K objects), ABO (2K articulated), and OmniObject3D [29] (6K objects) offer high fidelity but face scalability issues due to time-consuming scanning (15-60 mins/object).

Large-scale collections: Objaverse1.0 [7] (800K+ meshes) and ObjaverseXL [6] (10.2M meshes) provide scale but suffer from inconsistent quality and sparse coverage (e.g., only 127 mugs in Objaverse1.0).

GenVegeFruits3D [10] generates 3D assets but is limited to symmetric produce, avoiding challenges of arbitrary shapes. Its CPU-based texturing also increases overhead (see 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 from generative 3D model (SAI).

3D dataset	R/S	#Obj	#Cat	#O/C	Alignment	Quality	Time/mesh
ShapeNet [3]	S	51k	55	927	**yes**	**	X
ModelNet [30]	S	12k	40	300		**	X
3D-Future [12]	S	16k	34	470		**	X
ABO [5]	S	8k	63	159		**	X
Toy4K [22]	R	4k	105	38		**	15min-1h
GSO [9]	R	1k	17	59		****	15min-1h
AKB-48 [20]	R	2k	48	42		****	15min-1h
Omni3D [29]	R	6k	190	32		****	15min-1h
Objaverse1.0 [7]	R+S	800k	-	100	**yes**	***	X
ObjaverseXL [6]	R+S	**10.2M**	-	-	no	**	X
GenVegeFruits3D [10]	SAI	100K	**100**	**1000**	**yes**	****	15min
GenNocs3D (Ours)	SAI	6K	**6**	**1000**	**yes**	****	**3min**

As summarized in Table 1, our approach, *GenNOCS3D*, uniquely combines large-scale instance diversity—exceeding that of Objaverse1.0—with built-in canonical alignment and significantly faster generation times (5–20$\times$ speedup compared to traditional 3D scanning). This combination enables the efficient creation of high-quality, consistently aligned meshes that are directly usable for downstream 3D understanding tasks, without the limitations imposed by existing 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 [25] (6 categories) established the first benchmark, followed by PhoCAL [27], Wild6D [13], and HouseCat6D [18], which improved diversity but still suffer from sparse category coverage. Recent efforts like Omni6D [36] and Omni6Dpose [34] combine synthetic and real data but face two key issues: (1) limited instance diversity due to reliance on OmniObject3D [29] scans, and (2) reproducibility challenges from closed-source rendering pipelines.

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* which integrates two complementary rendering pipelines for sim2real transfer. We further release both pipelines along with

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

Cat6D dataset	3D dataset	Rendering	#Cat	#O/Cat	#Img
NOCS-CAMERA25 [26]	ShapeNet [3]	Rast+[26]	6	180.8	300K
NOCS-REAL275 [26]	3D scanned	R	6	7	8K
Phocal [28]	3D scanned	R	8	7.5	3.9K
Wild6D [13]	3D scanned	R	5	344,4	10K
HouseCad6D [18]	3D scanned	R	10	19	23.5K
Omni6DPose-SOPE [34]	Omni3D [29]	RT+MR+[26, 2, 33]	149	27.9	475K
Omni6DPose-ROPE [34]	3D scanned	R	149	3.8	332K
Omni6D [36]	[29]	RT+ [23]	166	28.2	0.8M
Omni6D-xl [36]	[29, 7, 6]	RT+ [23]	**419**	38.1	1.1M
Omni6D-real [36]	3D scanned	R	39	1.87	1K
GenNocs6D 4 (Ours)	GenNocs3D 3	RT+MR+[26, 23]	6	**1000**	2x300K

two large-scale datasets, each comprising 300K high-quality synthetic images, 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 category prompts. 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$\times$ compared to prior work [10].

3.1 Pipeline architecture

Our pipeline (illustrated by Fig. 1) achieves 3D generation from scratch with minimal human intervention: it requires only 2 hours of manual effort to produce the complete NOCS3D dataset. Our process comprises four phases:
1. Geometry prompt engineering: LLMs generate category-specific prompts with randomized shape variations, producing 100 initial images per category. Human filtering (<20 minutes/category) removes outliers while preserving geometric diversity and ensuring pose consistency.
2. Depth-conditioned generation: DepthFM [14] processes selected images (seconds per image) to create depth maps that condition subsequent generations, ensuring pose consistency.
3. Texture variation: Each depth map condition generates 10 texture-varied instances through additional LLM texture prompts, yielding 1000 total images per category. This enables maintaining image pose consistency and maximizing realistic shape and texture variations (see Fig. 1-2).
4. 3D reconstruction: A state-of-the-art Image-to-3D is applied to finally obtain consistently aligned 3D meshes (see Fig. 2), ready for 3D and 6D learning 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.

However, achieving reliable image-level pose consistency remains a major challenge. Our experimentation reveals 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 reasons behind this limitation: (1) current diffusion models lack an explicit understanding of 3D structure in the text-to-image generation process, and (2) natural language prompts inherently provide ambiguous pose specifications. While prompt engineering can occasionally improve pose consistency, it often introduces unrealistic visual artifacts and fails to resolve underlying ambiguity. To address these challenges, we adopt ControlNet [35] with depth-based conditioning. This approach enforces pose consistency during generation and achieves over 100%

success for simple symmetric objects, while dramatically improving pose consistency for complex shapes—reaching success rates as high as 98%.

Table 3: Image pose consistency rates (in %) across NOCS categories showing depth conditioning's impact (97.6% vs 57% baseline).

Method	Bottle	Bowl	Camera	Can	Laptop	Mug	Avg
Text-only [10]	70	70	30	70	20	82	57
Depth-conditioned (Ours)	100	100	97	100	90	100	96.5

As shown in Table 3, using depth-conditioned images increases the average pose consistency across NOCS categories from 57% to 97.6%. We also evaluated alternative conditioning methods, such as Canny edge detection, which often introduced visual artifacts due to missing or extraneous edges, and preserved unnecessary texture. In contrast, depth maps proved to be the most effective conditioning signal: they capture global object structure and spatial positioning without imposing constraints on local geometric variations or fine texture details. This enables the generation of images with texture details that are both pose-consistent and geometrically diverse, as illustrated in Fig. 2.

3.3 Comparison of image to 3D mesh generation

Recent advances in 3D mesh generation from single images 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 for grasping tasks; (iii) reliability across diverse object categories.

We benchmarked four recent methods that meet our time constraint. FS3D [1], SPAR3D [16], InstantMesh [32], and Hunyuan3D-v2.0 [24]. As shown in Table 4, our evaluation reveals that: (i) FS3D and SPAR3D offer the fastest generation ($<$10s), but produce inconsistent quality (particularly for concave objects like mugs); (ii) InstantMesh provides better quality at $<$1min generation time, but struggles with non-convex geometries critical for robotic grasping; (iii) Hunyuan3D-v2.0 achieves the best quality and achieve strong reliability in large scale generation avoiding manual mesh filtering.

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 [36] (Section 4.1) while also incorporating the mixed-reality rendering approach introduced in Omni6DPose [34] (Section 4.2). The two methods enable generation of a category-based 6D pose dataset in the same format with all the needed annotations: RGB-D, Masks, NOCS map[26], 6D poses.

4.1 Complete 3D simulation approach

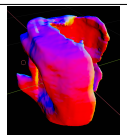
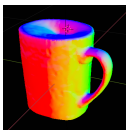
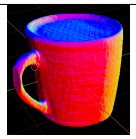
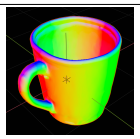
The first approach, adopted by Omni6D [36], uses full, realistic synthetic scenes from homes scanned in the real world. In this setup, the full-texture 3D scanned scene is loaded into the simulation. Next, we randomly placed objects in physically delineated areas and sampled 10 different camera viewpoints 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 [26] and enhanced by Omni6DPose [34]. 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 from the IKEA dataset (see Fig. 8) with the background image's viewpoint (as illustrated

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 [32]	Hunyuan3D-v2.0 [24]
Quality	⋆	⋆	⋆⋆	⋆⋆⋆
Time	<10s	<10s	<1m	<1m
Mesh				

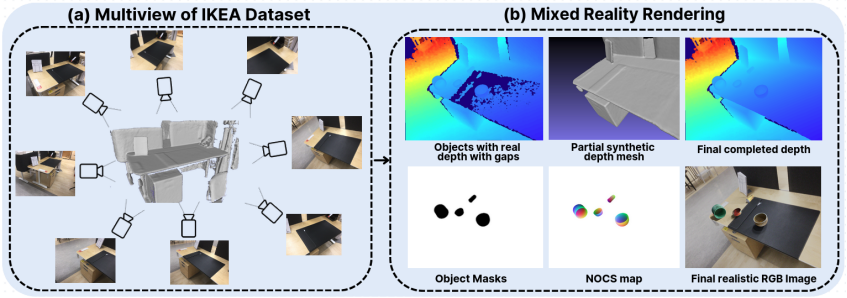


Figure 3: Mixed-reality data generation pipeline. (a) Simulation setup using IKEA dataset objects with registered camera poses for each background image. (b) Resulting RGB-D output showing: (top) RGB rendering with dynamic shadows, (middle) hybrid depth map combining real background and synthetic object depth, and (bottom) complete scene depth with integrated synthetic objects.

by Fig. 3(a)). Next, we placed between 4 and 8 objects per scene using camera ray casting to position them on planar surfaces. To ensure physically plausible poses, gravity was 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, as illustrated in Fig. 9. 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 as shown in Fig. 3(b). Finally, we used this pipeline to generate and release a dataset of 300,000 images corresponding to the original size of the NOCS dataset.

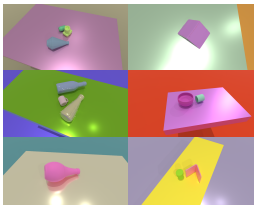


Figure 4: Images with random textures.

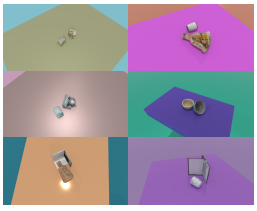


Figure 5: Images with object textures.

Figure 6: Real dataset of NOCS dataset: REAL275 [26].

Figure 7: Synthetic dataset of NOCS dataset: CAMERA25 [26].

Figure 8: REPLICA-based dataset (Ours).

Figure 9: Mixed-reality IKEA dataset (Ours).

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 [31], 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 integrating 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 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 collision, visual, and URDF files.

5.2 SAPIEN scene generation

For scene generation and training 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 estimation: NOCS baselines

This section presents different variations of the dataset generation to evaluates the Sim2Real performance of our pipeline. We evaluate the sim2real on the test part of Real275 the NOCS benchmark. For category-level 6D pose estimation, we have adopted DualPoseNet [19], which has demonstrated state-of-the-art performance on the Omni6D benchmark [36].

To systematically analyze different aspects of our method, we have generated 5 distinct dataset variations addressing the following research questions: (i) How does our mixed-reality setup (presented in Section 4.2) compare to fully synthetic approaches (presented in Section 4.1) in Sim2Real transfer? (ii) What is the performance impact of using our generated meshes compared to the original synthetic meshes in the NOCS dataset? (iii) How does the presence of shadows affect Sim2Real performance for both NOCS objects and our generated meshes?

Each variant of the data set contains approximately 300 K images, enabling a comprehensive evaluation across these dimensions. Table 5 highlights the best validation (val) scores in bold and demonstrates sim2real validation, further supporting the use of our generated objects as discussed in Section C.

Table 5: Evaluation of DualPoseNet on synthetic associated dataset and test real NOCS dataset.

Dataset	IoU_{50}	IoU_{75}	$5°2\,cm$	$5°5\,cm$	$10°2\,cm$	$10°5\,cm$	$5°$	$10°$	$2\,cm$	$5\,cm$	Avg
Replica-test	82.03	34.33	3.40	4.94	11.51	17.12	5.75	18.66	**55.24**	**98.05**	33.10
Mix$*SAI^{sh}$-test	**85.91**	**35.42**	4.62	**7.36**	**13.95**	**22.19**	**8.49**	**23.83**	49.29	96.43	**34.75**
Mix$*SAI^{no-sh}$-test	84.89	28.86	2.27	3.77	8.07	14.13	4.49	15.70	48.43	97.66	30.83
Mix$*syn^{sh}$-test	82.68	31.50	**4.80**	6.32	12.90	17.81	7.01	19.15	52.46	97.77	33.24
Mix$*syn^{no-sh}$-test	73.56	26.05	2.23	3.46	6.81	11.02	4.08	12.31	53.16	**98.05**	29.07
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$*SAI^{sh}$-val	**69.60**	**22.44**	**2.09**	**4.25**	**5.25**	**10.77**	**4.78**	**12.12**	26.31	81.47	**23.91**
Mix$*SAI^{no-sh}$-val	68.90	21.72	2.18	4.15	5.16	10.28	4.59	11.41	26.30	81.28	23.60
Mix$*syn^{sh}$-val	46.93	6.32	0.98	1.74	1.81	3.81	1.83	4.16	16.39	72.59	15.66
Mix$*syn^{no-sh}$-val	46.47	6.46	0.85	1.68	1.70	3.90	1.74	4.22	17.63	72.45	15.71

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 (↑). Configurations: `tex` = native textures, `rdom` = randomized textures, `val` = evaluation on 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**

6.2 Grasping and shape completion evaluation

We evaluated our approach within the CenterGrasp framework [4], using the SAPIEN simulator [31] 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, enabling a controlled analysis of texture impact 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 6, 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. Similarly to the original methods [4], we also demonstrate zero-shot transfer of the trained model in a real robot setup.

7 Limitation and future work

Although our proposed pipeline reliably generates large-scale, aligned 3D meshes, it currently requires minimal human filtering of depth inputs. Future work could focus on developing a fully autonomous pipeline by leveraging advances in controlled image processing [11] to directly infer 3D information for image generation.

Coverage and Generalization. Although our method is theoretically category-agnostic, we have currently validated it only on the six NOCS categories, which represent convex, non-convex, and holistic shapes. An important next step would be to expand evaluations to more diverse category variations, such as those benchmarked in [36, 34].

Computational Cost. Despite optimizations that reduced mesh generation time to approximately 3 minutes per object, large-scale dataset creation remains computationally demanding without GPU servers. While ongoing advances in 3D generation suggest this limitation may become less restrictive, future work should investigate more efficient generation techniques while maintaining rigorous quality standards.

6D Pose Estimation. Our demonstrated zero-shot sim2real transfer on the NOCS dataset confirms the viability of our generated data for 6D pose estimation tasks. However, extending this to more object categories, following the approach of [36], will be crucial for developing comprehensive benchmarks and ultimately enabling robust category-agnostic 3D/6D foundation models.

Robotic Grasping. While our aligned meshes successfully enable sim2real transfer in frameworks like CenterGrasp, we identify two promising research directions: (1) incorporating explicit category annotations to enhance training and better explore category-specific grasping strategies, and (2) evaluating grasping performance across a broader range of objects - though this expansion would require addressing the computational challenges of processing larger SDF-based mesh datasets.

8 Conclusion

In this paper, we have presented an end-to-end automated approach for generating high-quality 3D meshes, achieving a success rate of 97%—a significant improvement over existing pipelines (57%). Our method demonstrates robust performance across large-scale, non-convex objects while maintaining precise alignment.

To validate the utility of our generated meshes, we evaluated them in two key applications 6D Pose Estimation and grasping:

6D Pose Estimation: We showcased zero-shot sim2real transfer on the NOCS benchmark, achieving competitive performance without relying on any existing mesh datasets.

Robotics: We successfully integrated our meshes into a robotic simulator and conducted real-world experiments, confirming their applicability in grasping and manipulation tasks.

These results highlight that modern generative models are now mature enough to produce high-fidelity, aligned meshes for computer vision and robotics applications. By removing the dependency on manually created assets, our work unlocks new possibilities for the community to efficiently generate custom datasets tailored to specific needs—a critical step forward, as data scarcity often hinders progress in specialized domains.

Looking ahead, we believe generative models are a key pathway toward scalable 3D dataset creation, enabling the development of foundation models for 6D pose estimation and beyond. Future work will explore broader applications and further improvements in generalization.

References

[1] M. Boss, Z. Huang, A. Vasishta, and V. Jampani. Sf3d: Stable fast 3d mesh reconstruction with uv-unwrapping and illumination disentanglement. *arXiv preprint*, 2024.

[2] A. Chang, A. Dai, T. Funkhouser, M. Halber, M. Niessner, M. Savva, S. Song, A. Zeng, and Y. Zhang. Matterport3d: Learning from rgb-d data in indoor environments. *arXiv preprint arXiv:1709.06158*, 2017.

[3] A. X. Chang, T. Funkhouser, L. Guibas, P. Hanrahan, Q. Huang, Z. Li, S. Savarese, M. Savva, S. Song, H. Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015.

[4] E. Chisari, N. Heppert, T. Welschehold, W. Burgard, and A. Valada. Centergrasp: Object-aware implicit representation learning for simultaneous shape reconstruction and 6-dof grasp estimation. *IEEE Robotics and Automation Letters*, 2024.

[5] J. Collins, S. Goel, K. Deng, A. Luthra, L. Xu, E. Gundogdu, X. Zhang, T. F. Yago Vicente, T. Dideriksen, H. Arora, M. Guillaumin, and J. Malik. Abo: Dataset and benchmarks for real-world 3d object understanding. *CVPR*, 2022.

[6] M. Deitke, R. Liu, M. Wallingford, H. Ngo, O. Michel, A. Kusupati, A. Fan, C. Laforte, V. Voleti, S. Y. Gadre, et al. Objaverse-xl: A universe of 10m+ 3d objects. *Advances in Neural Information Processing Systems*, 36, 2024.

[7] M. Deitke, D. Schwenk, J. Salvador, L. Weihs, O. Michel, E. VanderBilt, L. Schmidt, K. Ehsani, A. Kembhavi, and A. 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.

[8] M. Denninger, D. Winkelbauer, M. Sundermeyer, W. Boerdijk, M. Knauer, K. H. Strobl, M. Humt, and R. Triebel. Blenderproc2: A procedural pipeline for photorealistic rendering. *Journal of Open Source Software*, 8(82):4901, 2023.

[9] L. Downs, A. Francis, N. Koenig, B. Kinman, R. Hickman, K. Reymann, T. B. McHugh, and V. 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.

[10] G. Duret, Y. Bourenname, D. Mazurak, A. Samsonenko, F. Zara, L. Chen, and J. Peters. Facilitate and scale up the creation of 3d meshes and 6d category-based datasets with generative models: Genvegefruits3d. *HAL*, 2025.

[11] A. Eldesokey and P. Wonka. Build-a-scene: Interactive 3d layout control for diffusion-based image generation. In *The Thirteenth International Conference on Learning Representations*, 2025.

[12] H. Fu, R. Jia, L. Gao, M. Gong, B. Zhao, S. Maybank, and D. Tao. 3d-future: 3d furniture shape with texture. *International Journal of Computer Vision*, pages 1–25, 2021.

[13] Y. Fu and X. 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.

[14] M. Gui, J. Schusterbauer, U. Prestel, P. Ma, D. Kotovenko, O. Grebenkova, S. A. Baumann, V. T. Hu, and B. Ommer. Depthfm: Fast generative monocular depth estimation with flow matching. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 3203–3211, 2025.

[15] T. Hodan, M. Sundermeyer, Y. Labbe, V. N. Nguyen, G. Wang, E. Brachmann, B. Drost, V. Lepetit, C. Rother, and J. 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.

[16] Z. Huang, M. Boss, A. Vasishta, J. M. Rehg, and V. Jampani. Spar3d: Stable point-aware reconstruction of 3d objects from single images. *arXiv preprint arXiv:2501.04689*, 2025.

[17] Z. Jiang, Y. Zhu, M. Svetlik, K. Fang, and Y. Zhu. Synergies between affordance and geometry: 6-dof grasp detection via implicit representations. *arXiv preprint arXiv:2104.01542*, 2021.

[18] H. Jung, S.-C. Wu, P. Ruhkamp, G. Zhai, H. Schieber, G. Rizzoli, P. Wang, H. Zhao, L. Garattoni, S. 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.

[19] J. Lin, Z. Wei, Z. Li, S. Xu, K. Jia, and Y. 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.

[20] L. Liu, W. Xu, H. Fu, S. Qian, Q. Yu, Y. Han, and C. 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.

[21] K. Mamou and F. 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.

[22] S. Stojanov, A. Thai, and J. M. Rehg. Using shape to categorize: Low-shot learning with an explicit shape bias. *CVPR*, 2021.

[23] J. Straub, T. Whelan, L. Ma, Y. Chen, E. Wijmans, S. Green, J. J. Engel, R. Mur-Artal, C. Ren, S. Verma, A. Clarkson, M. Yan, B. Budge, Y. Yan, X. Pan, J. Yon, Y. Zou, K. Leon, N. Carter, J. Briales, T. Gillingham, E. Mueggler, L. Pesqueira, M. Savva, D. Batra, H. M. Strasdat, R. D. Nardi, M. Goesele, S. Lovegrove, and R. Newcombe. The Replica dataset: A digital replica of indoor spaces. *arXiv preprint arXiv:1906.05797*, 2019.

[24] T. H. Team. Hunyuan3d 2.0: Scaling diffusion models for high resolution textured 3d assets generation, 2025.

[25] H. Wang, S. Sridhar, J. Huang, J. Valentin, S. Song, and L. 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)*, June 2019.

[26] H. Wang, S. Sridhar, J. Huang, J. Valentin, S. Song, and L. 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)*, June 2019.

[27] P. Wang, H. Jung, Y. Li, S. Shen, R. P. Srikanth, L. Garattoni, S. Meier, N. Navab, and B. Busam. PhoCaL: A Multi-Modal Dataset for Category-Level Object Pose Estimation with Photometrically Challenging Objects. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2022-June, pages 21190–21199. IEEE, jun 2022.

[28] P. Wang, H. Jung, Y. Li, S. Shen, R. P. Srikanth, L. Garattoni, S. Meier, N. Navab, and B. 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.

[29] T. Wu, J. Zhang, X. Fu, Y. Wang, L. P. Jiawei Ren, W. Wu, L. Yang, J. Wang, C. Qian, D. Lin, and Z. 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.

[30] Z. Wu, S. Song, A. Khosla, F. Yu, L. Zhang, X. Tang, and J. 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.

[31] F. Xiang, Y. Qin, K. Mo, Y. Xia, H. Zhu, F. Liu, M. Liu, H. Jiang, Y. Yuan, H. Wang, L. Yi, A. X. Chang, L. J. Guibas, and H. Su. SAPIEN: A SimulAted Part-Based Interactive ENvironment. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11094–11104. IEEE, jun 2020.

[32] J. Xu, W. Cheng, Y. Gao, X. Wang, S. Gao, and Y. Shan. Instantmesh: Efficient 3d mesh generation from a single image with sparse-view large reconstruction models. *arXiv preprint arXiv:2404.07191*, 2024.

[33] C. Yeshwanth, Y.-C. Liu, M. Nießner, and A. Dai. Scannet++: A high-fidelity dataset of 3d indoor scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12–22, 2023.

[34] J. Zhang, W. Huang, B. Peng, M. Wu, F. Hu, Z. Chen, B. Zhao, and H. 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.

[35] L. Zhang, A. Rao, and M. 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.

[36] M. Zhang, T. Wu, T. Wang, T. Wang, Z. Liu, and D. 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.

A 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 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.

B Code implementation use case

This section outlines the implementation of our dataset generation pipeline for the generation of any objects. For full implementation details, please refer to our GitHub repository: `https://github.com/Guillaume-Duret/GenNOCS`.

While we primarily evaluated our pipeline using the six categories from the NOCS dataset (to enable direct benchmarking with established methods), the system is designed to handle arbitrary object categories - including complex and non-convex geometries. Although comprehensive testing across numerous categories remains future work, the NOCS objects represent significant shape diversity, demonstrating the method's robustness. Importantly, we have encountered no theoretical limitations regarding shape complexity in our framework.

C Analyse Category 6D pose estaimtion and Sim2real comparison of our dataset on val set and real275 dataset

Table 7: Evaluation of DualPoseNet on synthetic associated dataset and test real NOCS dataset.

Dataset	IoU_{50}	IoU_{75}	$5°2\,cm$	$5°5\,cm$	$10°2\,cm$	$10°5\,cm$	$5°$	$10°$	$2\,cm$	$5\,cm$	Avg
Replica-test	82.03	34.33	3.40	4.94	11.51	17.12	5.75	18.66	**55.24**	**98.05**	33.10
Mix$*SAI^{sh}$-test	**85.91**	**35.42**	4.62	**7.36**	**13.95**	**22.19**	**8.49**	**23.83**	49.29	96.43	**34.75**
Mix$*SAI^{no-sh}$-test	84.89	28.86	2.27	3.77	8.07	14.13	4.49	15.70	48.43	97.66	30.83
Mix$*syn^{sh}$-test	82.68	31.50	**4.80**	6.32	12.90	17.81	7.01	19.15	52.46	97.77	33.24
Mix$*syn^{no-sh}$-test	73.56	26.05	2.23	3.46	6.81	11.02	4.08	12.31	53.16	**98.05**	29.07
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$*SAI^{sh}$-val	**69.60**	**22.44**	**2.09**	**4.25**	**5.25**	**10.77**	**4.78**	**12.12**	26.31	81.47	**23.91**
Mix$*SAI^{no-sh}$-val	68.90	21.72	2.18	4.15	5.16	10.28	4.59	11.41	26.30	81.28	23.60
Mix$*syn^{sh}$-val	46.93	6.32	0.98	1.74	1.81	3.81	1.83	4.16	16.39	72.59	15.66
Mix$*syn^{no-sh}$-val	46.47	6.46	0.85	1.68	1.70	3.90	1.74	4.22	17.63	72.45	15.71

Table 7 presents the ablation study results comparing different dataset generation approaches for 6D pose estimation and sim2real performance. The validation set evaluates generalization to new scenes while maintaining training-like conditions to assess category-level 6D pose learning, while the test set provides zero-shot evaluation on the real-world NOCS REAL275 benchmark [26].

In the validation results, the comparison between MixSyn (synthetic objects) and MixSai (our generated objects) reveals a significant improvement in category-level 6D pose estimation, with average performance increasing from 15.69 to 23.7. This enhancement stems from our objects'

superior category consistency compared to synthetic alternatives, which often exhibit problematic mesh variations - for instance, TVs being misclassified as laptops or full-screen displays appearing in the laptop category. Notably, shadow effects show minimal impact on sim2sim evaluation, while our shadow-based mixed reality setup achieves the best overall performance.

For sim2real transfer (test results), the mixed reality with shadows configuration demonstrates the strongest performance with an average score of 34.75, validating our pipeline's effectiveness. The results particularly highlight the importance of proper shadow configuration, which proves critical for sim2real performance across both synthetic and generated objects. These findings underscore how our approach overcomes the limitations of synthetic datasets while maintaining robust real-world applicability.

The overall results demonstrate the reliability of our pipeline in generating Category 6D pose datasets from scratch while effectively reducing the sim2real gap. Figure 10 shows pose estimation examples from the validation set for both the Mixed Reality and Replica datasets.

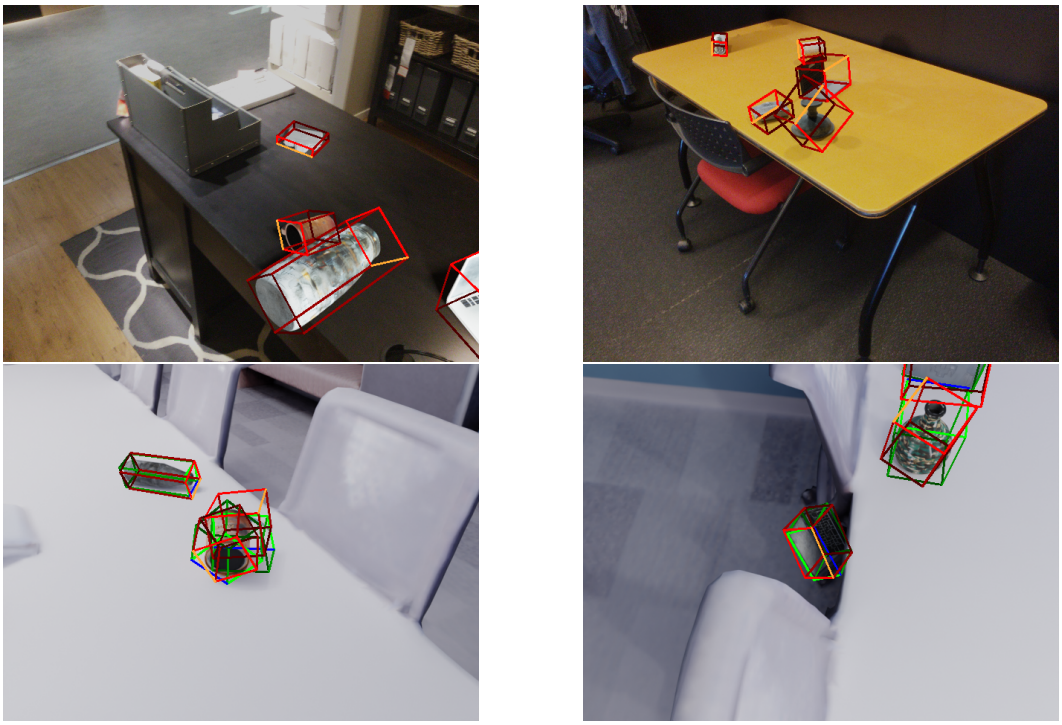


Figure 10: Inference example of our model on the validation set.

D Real robot grasping experiment

For the grasping task, we demonstrate zero-shot transfer of our trained model using a 7-joint Franka Panda arm equipped with a ZED Mini camera. Figure 11 shows examples of generated grasping poses in real-world cluttered scenes.

E 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, REPLICA dataset, and the two grasping datasets.

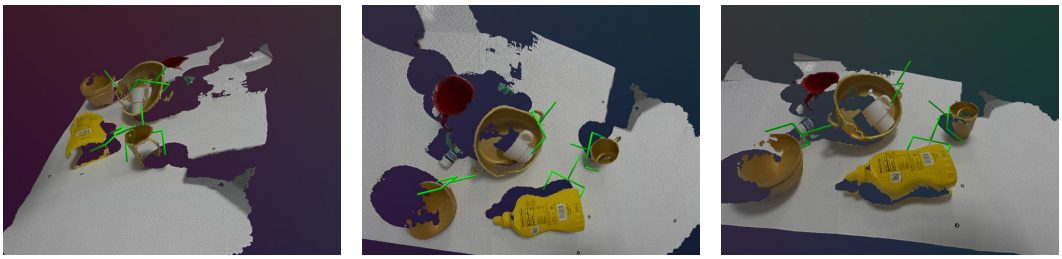


Figure 11: Grapsping generation on a real-world scene.

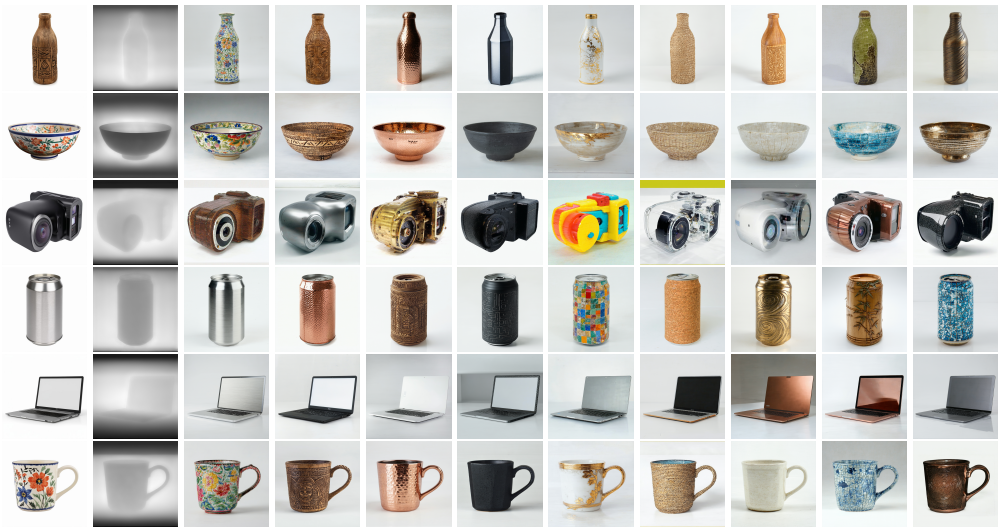


Figure 12: Example of images generated through depth conditioning, the first image is initial image, the second is the depth, the following images are conditioned based textured images with different texture prompts.

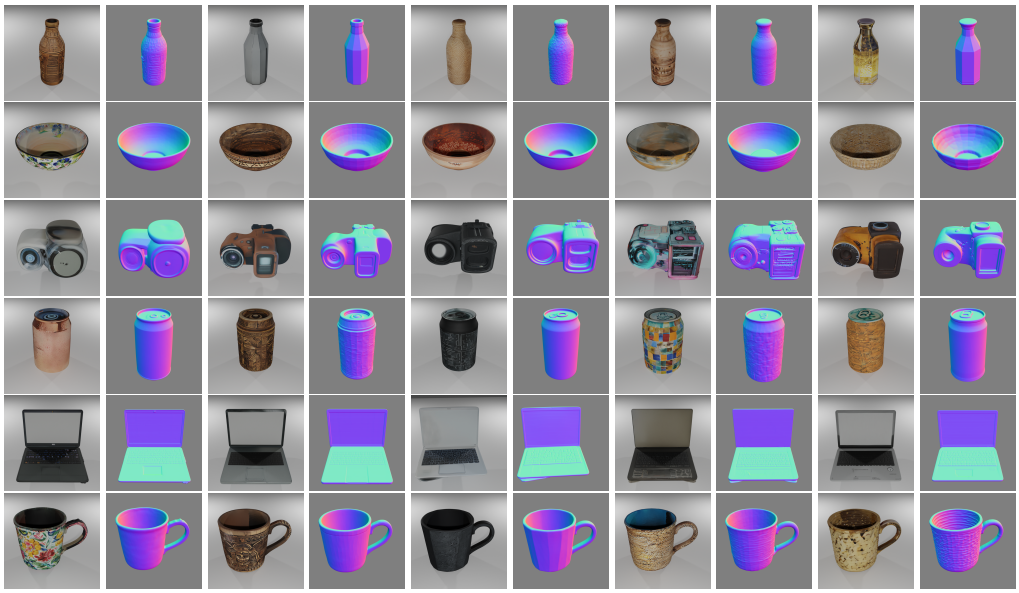


Figure 13: Different input features generated from prompt-based generated images.

Figure 14: 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 15: 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.

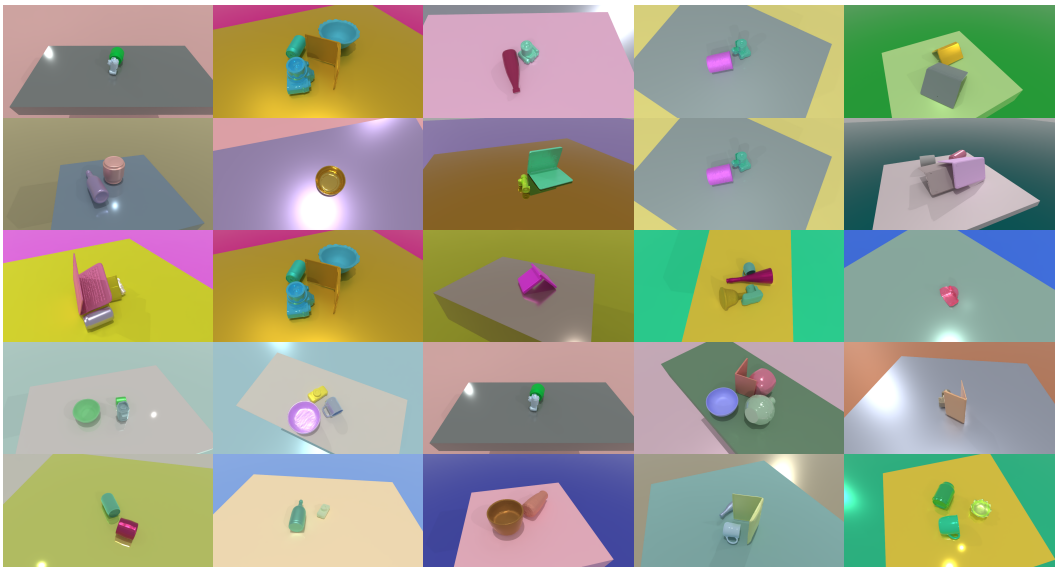


Figure 16: Examples of random grasp attempts on various objects. Each row shows different object categories and grasp approaches.

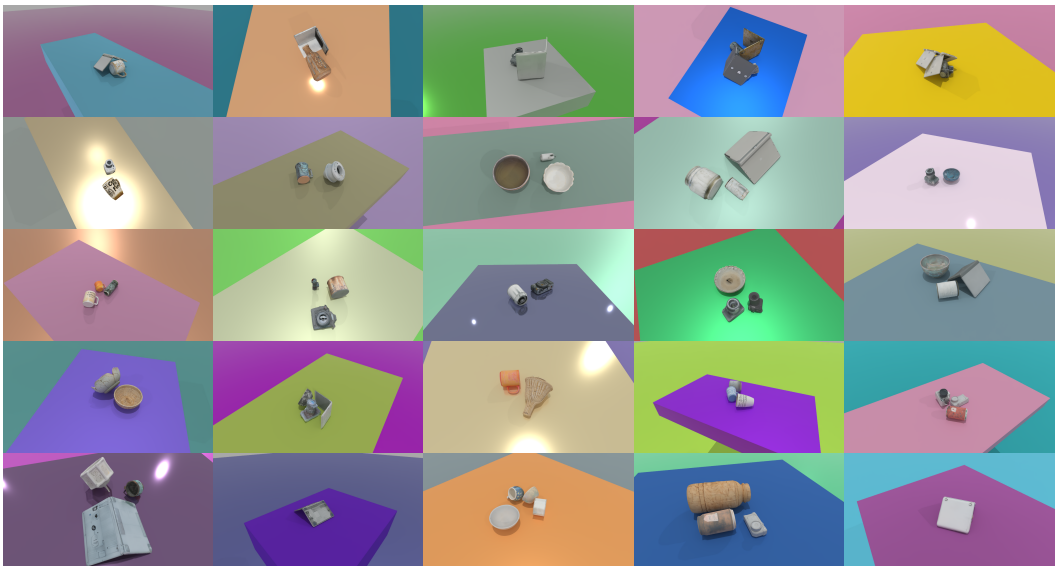


Figure 17: Texture-based grasp attempts on various vegetables and produce. The images demonstrate different approach angles and contact points.