IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

GraspClutter6D: A Large-scale Real-world Dataset for Robust Perception and Grasping in Cluttered Scenes

Seunghyeok Back[1], Joosoon Lee[2], Kangmin Kim[2], Heeseon Rho[2], Geonhyup Lee[2], Raeyoung Kang[2], Sangbeom Lee[2], Sangjun Noh[2], Youngjin Lee[2], Taeyeop Lee[3] and Kyoobin Lee[2†]

***Abstract*— Robust grasping in cluttered environments remains an open challenge in robotics. While benchmark datasets have significantly advanced deep learning methods, they mainly focus on simplistic scenes with light occlusion and insufficient diversity, limiting their applicability to practical scenarios. We present GraspClutter6D, a large-scale real-world grasping dataset featuring: (1) 1,000 highly cluttered scenes with dense arrangements (14.1 objects/scene, 62.6% occlusion), (2) comprehensive coverage across 200 objects in 75 environment configurations (bins, shelves, and tables) captured using four RGB-D cameras from multiple viewpoints, and (3) rich annotations including 736K 6D object poses and 9.3B feasible robotic grasps for 52K RGB-D images. We benchmark state-of-the-art segmentation, object pose estimation, and grasp detection methods to provide key insights into challenges in cluttered environments. Additionally, we validate the dataset's effectiveness as a training resource, demonstrating that grasping networks trained on GraspClutter6D significantly outperform those trained on existing datasets in both simulation and real-world experiments. The dataset, toolkit, and annotation tools are publicly available on our project website: `https://sites.google.com/view/graspclutter6d`.**

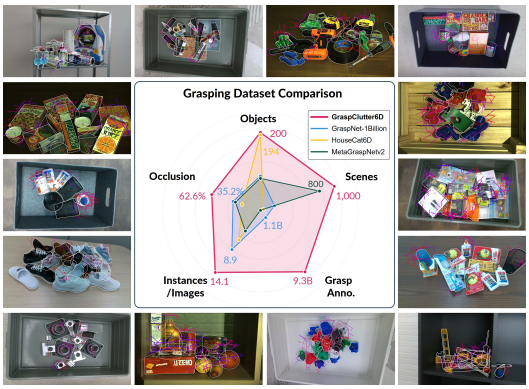

Fig. 1: We present **GraspClutter6D**, a large-scale dataset with high complexity (avg. **14.1** instances, **62.6%** occlusion), including **200** objects in **1,000** scenes from four RGB-D sensors. It offers **736K** 6D object poses and **9.3B** grasps. (25 grasps are visualized per image (avg. 178K grasps / image))

I. INTRODUCTION

Grasping is one of the most fundamental yet challenging tasks in robotics, with applications spanning warehouses, manufacturing, and household assistance. Effective robotic grasping requires coordination of multiple visual perception capabilities: segmentation [1]–[3], 6D object pose estimation [4]–[6], and 6-DoF grasp detection [7]–[9]. Recent significant progress has been driven by comprehensive datasets [4], [7], [9]–[15], enabling deep learning to achieve better generalization. However, grasping in *cluttered environments*—a common scenario in practical settings—remains challenging. In these environments, objects are densely packed in unknown poses under varied backgrounds, requiring sophisticated approaches to handle occlusion [16], [17].

Although datasets have advanced this field, most existing benchmarks focus on structured, simplified scenes rather than cluttered environments. GraspNet-1Billion (GraspNet-1B) [7], [18], a widely adopted benchmark, exhibits only modest complexity (avg. 8.9 objects per scene with 35.2% occlusion) and restricts diversity to single tabletop scenes without background variation. Similarly, the recent HouseCat6D dataset [19] offers greater object diversity (192 objects) but limited scene composition, containing only 41 scenes with minimal occlusion (23.5%). This gap between simplified datasets and real-world complexity presents a significant challenge in developing robust robot manipulation systems.

In this work, we present **GraspClutter6D**, a large-scale dataset for robotic grasping and perception (Fig. 1), the most extensive real-world resource with diverse objects in highly cluttered environments. Through multi-sensor robotic capture and crowd-sourcing, we collected 1,000 highly cluttered scenes containing 9.3 billion grasp annotations and 736K instance annotations. Benchmark evaluations reveal that state-of-the-art grasping methods face significant challenges in our cluttered environments, while training on our dataset substantially improves grasping performance compared to existing datasets. We have released the complete dataset, annotation tools, and purchase links for all objects and furniture to facilitate research in robust robotic manipulation.

The contributions of this work are summarized as follows:

- **Real-world dense clutter**: 1,000 densely packed scenes with high complexity, averaging 14.1 instances with 62.6% occlusion, totaling 52K RGB-D images.
- **Diverse coverage**: Bin/shelf/table scenes with 200 objects captured using four cameras from multiple viewpoints.

This work was supported by the Technology Innovation Program (RS-2024-00442029, Development of Tactile Intelligence in Robotic Hands Based on Tactile Data Learning to Manipulate Irregular Multiple Types of Objects and RS-2024-00423940, Development of Humanoid Robots That Feel Like Humans, Communicate, and Grow through Learning) funded by the Ministry of Trade Industry & Energy(MOTIE, Korea).

[1] Korea Institute of Machinery & Materials (KIMM)
[2] Gwangju Institute of Science and Technology (GIST)
[3] Korea Advanced Institute of Science and Technology (KAIST)
S. Back was with GIST during initial research and is now with KIMM.
[†] Corresponding author: Kyoobin Lee `kyoobinlee@gist.ac.kr`.

TABLE I: **Comparison of real-world 6D pose estimation and grasping datasets.** GraspClutter6D provides highly complex scenes at scale in diverse environments. *Visibility*: mean visible pixel ratio per instance; *Occlusion*: percentage of instances with Visibility ≤ 0.95. *OCID: planar grasps only; others: 6D grasps. (K=thousands, M=millions, B=billions)

Dataset	Annotations			Scale					Scene Complexity			
	Segm.	6D Pose	Grasp	Objects	Scenes	Sensor Type(s)	Grasps	Images	Instances / Image ↑	Visibility (%) ↓	Occlusion (%) ↑	Environment Type(s)
Shelf&Tote [11]	✓	✓	✗	39	452	1	·	7K	4.6	-	-	bin, shelf
Rutgers APC [10]	✓	✓	✗	25	-	1	·	10K	<5	-	-	shelf
YCB-Video [4]	✓	✓	✗	21	92	1	·	130K	4.5	86.1	47.3	table
T-LESS [12]	✓	✓	✗	30	20	3	·	48K	∼7	-	-	table
MP6D [20]	✓	✓	✗	20	77	1	·	20K	6.2	-	-	table
PACE [21]	✓	✓	✗	238	300	1	·	55K	4.7	85.5	41.5	indoor, no shelf
OCID-Grasp [13], [22]	✓	✗	✓*	89	96	1	75K	2K	7.5	-	-	table, floor
MetaGraspNetv2 [14]	✓	✓	✓	82	800	1	-	32K	4.7	90.8	32.1	bin
HouseCat6D [19]	✓	✓	✓	194	41	2	10M	25K	6.7	96.0	23.5	table
GraspNet-1B [7], [18]	✓	✓	✓	88	190	2	1.1B	**97K**	8.9	90.9	35.2	table
GraspClutter6D (Ours)	✓	✓	✓	**200**	**1,000**	**4**	**9.3B**	52K	**14.1**	**77.1**	**62.6**	**bin, shelf, table**

- **Extensive annotations**: 9.3 billion 6-DoF grasp poses with 736K 6D object poses and segmentation masks.
- **Improved grasping and perception**: Models trained on our dataset demonstrate significantly better performance than those trained on existing benchmark datasets.
- **Benchmark evaluation**: Assessment of state-of-the-art grasping and perception methods to provide baselines.

II. RELATED WORK

This section reviews key methods and datasets in robotic perception and grasping. Table I summarizes real-world datasets for 6D object pose estimation and robotic grasping.

Grasping in Cluttered Scenes. Traditional parallel-jaw grasping [23] relied on known object models, registering CAD models to scene point clouds for predefined grasps. Learning-based methods [8], [9], [24]–[27] subsequently enabled the grasping of unknown objects. GPD [24] pioneered a deep network for 6-DoF grasps, while DexNet [28] facilitated planar grasping through large-scale learning on synthetic data. Recent grasping networks, such as Contact-GraspNet [8] and AnyGrasp [9], directly predict dense 6-DoF grasps from visual scenes by learning from large-scale datasets [7], [18], [29]. Nevertheless, robust grasping in complex, cluttered environments remains challenging, as simulation datasets [14], [29] suffer from sim-to-real transfer gaps, while existing real-world datasets [7], [18], [19] primarily feature simplified, structured scenes with low occlusion and limited diversity. We address these limitations by providing an extensive real-world grasping dataset featuring highly cluttered scenes across diverse environments and objects.

Segmentation and 6D Object Pose Datasets. Robotic perception primarily consists of segmentation [30], [31] to locate object instances and 6D pose estimation [5], [32] for precise manipulation. Early segmentation datasets, such as OSD [33] and OCID [13], provided foundational RGB-D scenes but were limited in scale, offering at most 2K images. ArmBench [34] includes 53K warehouse images but lacks the 6D pose data critical for robust manipulation. For 6D object pose estimation, the BOP benchmark [15] has advanced progress through standardized evaluations, featuring datasets such as YCB-Video [4], T-LESS [12], and Rutgers APC [10]. However, these datasets contain relatively lightly cluttered setups—averaging 4.5 [4] to 7 [12] objects per scene—with moderate occlusion (up to 47.3% in [4]). Similarly, the recent PACE dataset [21], despite encompassing 238 objects, maintains modest complexity with 4.7 objects per scene and 41.5% occlusion. In contrast, GraspClutter6D offers 736K object poses in densely packed scenes (averaging 14.1 objects, 62.6% occlusion) with 9.3B grasp annotations.

Robotic Grasping Datasets. State-of-the-art robotic grasping methods rely on deep learning approaches, which necessitate large-scale, diverse datasets for robust performance. Early datasets such as Cornell [35] and Jacquard [36] provided valuable resources for grasp learning. However, they were limited to planar grasping in single-object scenes. OCID-Grasp [22] extended this paradigm to cluttered scenes but offered only 75K planar grasp annotations. Synthetic datasets can be scalable alternatives: DexNet [25] with 6.7M planar grasps, ACRONYM [29] with 6-DoF grasps for 8K objects, and MetaGraspNetv2 [14] with 8,000 synthetic scenes. However, they face significant sim-to-real gaps, such as inaccurate sensor noise and contact dynamics. Among real-world 6-DoF grasp datasets, only a few public datasets exist. GraspNet-1B [7], [18], a widely adopted grasping benchmark, provides 1.1B grasp poses across 190 scenes. Though it established a foundation for 6-DoF grasping research, it remains confined to simple green tabletop environments with relatively low complexity (35.2% occlusion) and no background variation. The recent HouseCat6D dataset [19] enhances diversity with 194 objects, yet remains limited in scale and variation, offering only 10M grasps for 41 table-only scenes with low occlusion (23.5%). GraspClutter6D bridges these gaps by providing 9.3B 6-DoF grasp annotations across 1,000 highly cluttered scenes.

III. GRASPCLUTTER6D DATASET

This section presents our systematic approach to establish a comprehensive dataset for robotic manipulation in complex environments. We detail the data acquisition and annotation process, followed by key statistics and quality evaluation.

A. Data Acquisition

3D Object Models. Our dataset comprises 200 objects (Fig. 2), carefully curated to span household, warehouse,

Fig. 2: **Diverse 3D object models from GraspClutter6D**. The numbers in brackets denote objects per category.

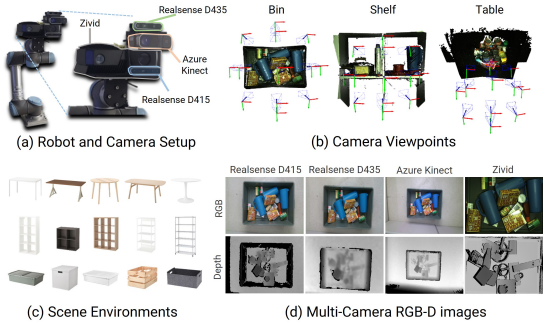

Fig. 3: **Multi-camera capture system:** (a) a UR5 with four cameras, (b) multiple camera viewpoints, (c) bin, shelf, and table setup, (d) RGB-D image comparison.

and industrial domains. We scanned 108 custom objects and incorporated 92 items from established benchmarks—YCB [37], HOPE [38], GraspNet-1B [7], [18], DexNet [25], and APC [39]—to ensure compatibility. High-quality 3D textured models were generated for custom objects using an Artec Leo scanner followed by post-processing by experts. This process yielded watertight, hole-free meshes with detailed textures. Reflective and transparent objects, which typically cause scanning artifacts, were coated with non-reflective gray spray to capture accurate geometry. The resulting 3D models are available in unified formats, supporting annotation and potential synthetic data generation. We also provide purchase links for all objects for future robotics research.

Hardware Setup. We developed a multi-sensor robotic capture system for efficient multi-viewpoint scene collection (Fig. 3 (a)-(c)). Four RGB-D cameras were mounted on a UR5 robot arm using a rigid rig: three low-cost commercial sensors (RealSense D415, D435, and Azure Kinect) and one sub-millimeter high-precision sensor (Zivid One+ M). These cameras were synchronized to capture simultaneous data. To create diverse yet reproducible environments, we selected 5 IKEA furniture pieces from each of three categories (tables, shelves, and bins) with 5 background variations per category, yielding 75 unique configurations.

Camera Calibration. We performed extensive calibration for high-quality RGB-D data. We first conducted intrinsic calibration using a ChArUco board [40]. Then, relative poses

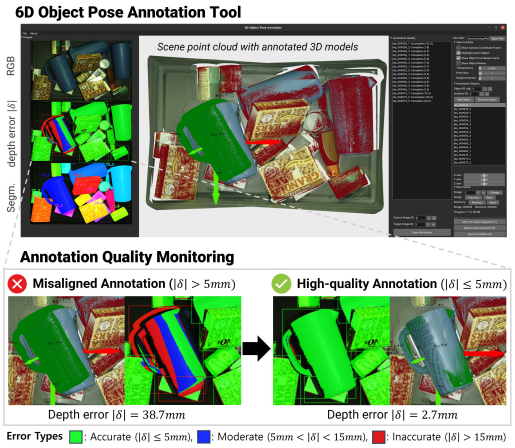

Fig. 4: **Pose annotation tool with quality monitoring.** Top: The interface displays point cloud, RGB, and annotation quality. Bottom: Comparison between misaligned and high-quality annotation. Color-coded error visualization guides annotators to achieve the target accuracy threshold of 5 mm.

between cameras were obtained using MC-Calib [41], followed by bundle adjustment with PnP-RANSAC [42] applied to matched points [43], [44] on key frames. Camera poses relative to the robot were determined using ArUco markers [40]. To address systematic depth errors in low-cost RGB-D cameras, we applied depth correction through least-squares fitting with linear models [12]. This reduced mean absolute depth errors from 10.2 mm to 5.0 mm for the RealSense D415, from 15.6 mm to 7.0 mm for the RealSense D435, and from 25.4 mm to 15.8 mm for the Azure Kinect. Our dataset provides undistorted, depth-corrected RGB-D pairs with camera poses, requiring no additional post-processing.

Scene Capture. We generated 1,000 cluttered scenes by randomly placing 5-20 objects in each configuration. Each scene was recorded from 13 viewpoints (Fig. 3 (b))—one centered and twelve peripheral angles—ensuring comprehensive visual coverage. This process yielded 52,000 RGB-D images across all cameras. Example RGB-D images from different cameras are shown in Fig. 3 (d), capturing diverse fields of view, illuminations, and depth characteristics. Overall, the dataset features high scene complexity with an average of 14.1 object instances per image and 77.1% visibility, providing multi-camera, multi-view real-world data for robotic manipulation in cluttered environments.

B. Annotation Pipelines

Object Pose Annotation. GraspClutter6D is annotated through crowd-sourcing with a custom annotation tool (Fig. 4). To ensure high-quality and efficient annotation, we used two strategies: 1) a quality monitoring system providing immediate visualization of annotation errors, and 2) annotation propagation, wherein each scene is annotated once on an integrated point cloud and then propagated to all views.

We first generated integrated point clouds of each scene by merging data from 13 different Zivid camera views.

Annotators then aligned 3D object models to these sub-millimeter accurate point clouds, targeting a mean depth error below $5\,\mathrm{mm}$ per object. Our tool visualized depth errors between annotated object models and scene point clouds, allowing annotators to monitor quality and refine alignments through additional adjustment and ICP [45] refinement upon saving each annotation. Independent experts verified all annotations through a double-review process. The verified poses were propagated to all camera views using camera intrinsic parameters and extrinsic poses. This process yielded $735,545$ annotated instances with 6D object poses and segmentation masks. The annotation tool is available on our project website.

Grasp Pose Annotation. We adopted a two-stage pipeline based on well-established methods in [7], [18] for 6-DoF grasp annotation: 1) object-level annotation with the force-closure metric, and 2) scene-level annotation with grasp projection and collision detection. We uniformly downsampled 200 object models in voxel space, sampling 14.4K grasps per point. For each grasp, the force closure metrics [25], [46] were calculated by varying the friction coefficients from 1.0 to 0.1. Then, object-level grasps were projected to scene-level grasps using annotated 6D object poses. Finally, collision checking was performed by filtering invalid grasps that overlap with the reconstructed scene point cloud. This process yielded 9.3B collision-free 6-DoF grasp poses for a parallel-jaw gripper (avg. 178K grasps per image), providing the largest grasp annotations in real cluttered scenes.

TABLE II: **Comparison of 6D object pose annotation accuracy across datasets**, measured by depth differences (δ) between captured and rendered depth at annotated poses. ($\mu_{|\delta|}$ and $med_{|\delta|}$: mean and median absolute depth differences, σ_δ: standard deviation of depth differences in mm.)

Dataset	Sensor	$\mu_{\lvert\delta\rvert}$	$med_{\lvert\delta\rvert}$	σ_δ
T-LESS [12]	Camarine	4.28	2.46	7.72
	Kinect v2	8.40	5.45	11.36
LINEMOD [47]	Kinect v2	5.89	5.57	1.47
YCB-Video [4]	Xtion Pro Live	3.95	3.66	2.26
MP6D [20]	Tuyang FM851-E2	3.54	2.70	0.17
GraspNet-1B [7]	RealSense D435	7.69	4.95	14.30
	Azure Kinect	14.79	9.54	20.20
GraspClutter6D (Ours)	Zivid	**3.22**	**1.55**	11.10
	RealSense D415	5.71	3.77	14.67
	RealSense D435	7.02	4.58	13.82
	Azure Kinect	13.85	6.83	32.59

C. Dataset Statistics

Annotation Accuracy. The annotation quality of 6D object poses was measured using the depth difference ($\delta = d_r - d_c$) as proposed by [12]. This metric calculates the difference between rendered depth (d_r) based on annotated object poses and captured depth (d_c) acquired from sensors for valid depth pixels (non-zero and finite) in both images. Values exceeding $5\,\mathrm{cm}$ were excluded following the standards to remove outliers from occlusion. Thus, δ quantifies pose annotation errors with sensor depth precision.

Table II presents the comparison of annotation accuracy with existing datasets. The mean and median absolute depth difference ($\mu_{|\delta|}$ and $med_{|\delta|}$) measure error magnitude, while standard deviation (σ_δ) indicates consistency of annotation and sensor measurement. For a fair comparison with other datasets, the values of GraspClutter6D were calculated on opaque objects [38], as reflective and transparent objects in our dataset inherently produce measurement artifacts in depth sensors. The results show that GraspClutter6D achieved high annotation accuracy with $\mu_{|\delta|}$ of $3.22\,\mathrm{mm}$ for the Zivid sensor. Compared to GraspNet-1B [7], [18], GraspClutter6D yielded higher accuracy with $\mu_{|\delta|}$ values of $7.02\,\mathrm{mm}$ and $13.85\,\mathrm{mm}$ for RealSense D435 and Azure Kinect, respectively, compared to their reported $7.69\,\mathrm{mm}$ and $14.79\,\mathrm{mm}$.

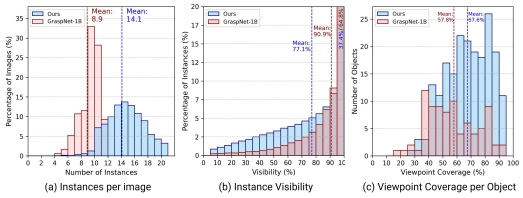

Fig. 5: **GraspClutter6D statistics compared to GraspNet-1B**, showing comprehensive coverage of our dataset.

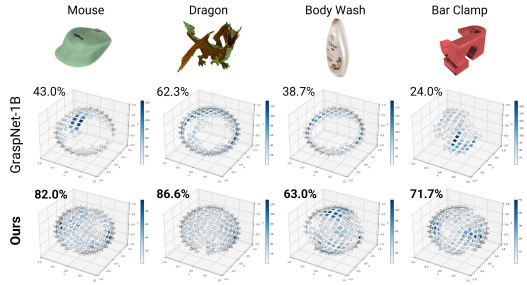

Fig. 6: **Viewpoint coverage comparison** between GraspNet-1B (top) and GraspClutter6D (bottom). The heatmaps represent object-to-camera orientations discretized into 300 uniform spherical bins. Darker blue indicates high density.

Scene Statistics. Our dataset comprises 1,000 real scenes featuring 200 objects with 735,545 instances annotated with 6D poses and segmentation masks, and 9.3 billion collision-free 6-DoF grasp poses. As shown in Fig. 5 (a)-(c), our dataset includes highly cluttered scenes (avg. 14.1 instances and 77.1% visibility), offering greater scene complexity than the GraspNet-1B (avg. 8.9 instances with 90.9% visibility). Object-to-camera distances range from $0.31\,\mathrm{m}$ to $1.37\,\mathrm{m}$ (average $0.77\,\mathrm{m}$), with object diameters between $4.5\,\mathrm{cm}$ and $47.3\,\mathrm{cm}$ and weights from $6\,\mathrm{g}$ to $2250\,\mathrm{g}$ (mean $199.5\,\mathrm{g}$)—dimensions suitable for common robotic grippers.

Viewpoint Coverage. Our dataset provides diverse viewpoint coverage through multi-view capture in bin, shelf, and table environments. We quantify this coverage by discretizing the object-to-camera rotations in spherical space into 300 uniform bins. As shown in Fig. 5 (d), the mean viewpoint coverage for all objects in GraspClutter6D is 67.6%, higher than 57.8% for GraspNet-1B. We also compared the viewpoint coverages for the 32 objects that overlap between the two datasets in Fig. 6. GraspClutter6D achieves a mean

viewpoint coverage of 73.8% for these overlapped objects, significantly higher than 48.5% of GraspNet-1B, due to its larger scale and more diverse capture environments.

Dataset Splits. For standardized evaluation, we provide two main splits: (1) A cross-object setup that tests generalization to novel objects. This split uses 68 unseen YCB-HOPE objects [37], [38] for testing (12K images, 235 scenes), widely used robotic research benchmark objects, and 132 objects for training (32K images, 413 scenes). (2) An intra-object setup focused on 21 common YCB-Video objects [4] used in pose estimation benchmarks. This configuration includes 19K training images from 370 scenes and 4K test images from 74 scenes. The dataset also provides metadata for researchers to create custom splits for specific tasks.

IV. Experiments

We conducted experiments with two primary objectives: (1) to evaluate the effectiveness of our dataset as a training resource for grasp detection and perception models, (2) to benchmark state-of-the-art methods on GraspClutter6D for future research. The following sections provide a comparative analysis of grasping performance across different training datasets, followed by benchmarks for segmentation, 6D object pose estimation, and 6-DoF grasp detection.

A. Impact of GraspClutter6D on 6-DoF Grasp Detection

Experimental Design. We investigated whether training on GraspClutter6D improves 6-DoF grasp detection performance. Using Contact-GraspNet [8] as our baseline network, we conducted comparative experiments in both simulated and real-world environments with three distinct datasets:

- *ACRONYM* [29]: A widely-used synthetic dataset with grasp annotations for 8,872 objects. Following protocols in [8], we generated synthetic scenes by randomly placing 8-12 objects in stable, non-colliding poses on a table surface.
- *GraspNet-1B* [7], [18]: A popular real-world dataset whose training set contains 100 table scenes featuring 30 objects arranged in moderately cluttered configurations.
- *GraspClutter6D*: Our dataset comprising 413 highly cluttered scenes across table, shelf, and bin environments with 132 objects in the training set (cross-object setup).

We also benchmarked against *AnyGrasp* [9], a state-of-the-art commercial grasp detection SDK trained on *GraspNet-1B++*, a non-public dataset that extends GraspNet-1B with extra objects and scenes (144 objects across 268 real-world scenes). We used AnyGrasp with its default configuration. Contact-GraspNet was trained for 30K iterations with a batch size of 80 for each dataset.

Simulation Setup. We used PyBullet [48] simulation from [49] with two configurations: 1) *packed* scenes, where 5 objects were placed in an upright pose without occlusion to test grasping in simple scenarios, and 2) *pile* scenes with randomly dropped objects with increased complexity (5, 10, and 15 objects) to evaluate performance under varying occlusion levels. For each setup, we generated random scenes and executed the highest-confidence grasp using a Panda gripper. Rounds continued until all objects were cleared, no valid grasps remained, or two consecutive grasps failed. We conducted 500 simulation rounds per method and measured performance using standard metrics [49]: Grasp Success Rate (GSR)—the ratio of successful lifts without slippage—and Declutter Rate (DR)—the average ratio of objects cleared per round. Objects out of workspace during manipulation were regarded as failures in DR. Single-view point clouds with Gaussian noise were used as input for all models.

TABLE III: **Simulated grasping results.** We report grasping success rate (GSR) and declutter rates (DR).

Method	Train Data	Packed		Pile					
		5 objects		5 objects		10 objects		15 objects	
		GSR	DR	GSR	DR	GSR	DR	GSR	DR
AnyGrasp	GraspNet-1B++ [9]	77.5	78.9	58.6	59.3	68.2	53.0	71.6	49.5
Contact-GraspNet	ACRONYM [29]	64.6	60.1	39.2	28.8	44.8	21.2	49.3	18.7
	GraspNet-1B [18]	74.5	65.2	62.7	46.6	67.9	36.3	71.0	33.4
	GraspClutter6D	**84.9**	**86.1**	**77.6**	**75.4**	**77.2**	**64.7**	**77.3**	**54.0**

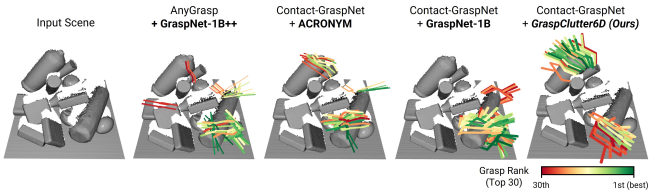

Fig. 7: Grasp predictions (top 30) in a **simulated pile scene**.

Simulation Results. Table III and Fig. 7 compare grasping performance across different training datasets. The results demonstrate that Contact-GraspNet trained on our GraspClutter6D dataset significantly outperformed models trained on ACRONYM and GraspNet-1B across all test scenarios. Notably, our model achieved Grasp Success Rates (GSR) of 84.9% in 5-object packed scenes and 77.3% in more challenging 15-object pile scenarios, showing better performance than AnyGrasp. This suggests that the complexity and diversity of GraspClutter6D make it particularly effective for training robust grasping systems in cluttered environments.

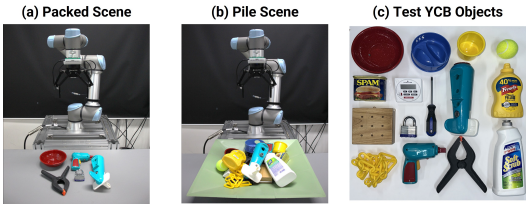

Fig. 8: Real-world robotic grasping setup and test objects.

Real-world Setup. We validated the effectiveness of GraspClutter6D in real-world robotic grasping. Similar to the simulation setup, we employed two configurations (Fig. 8): 1) *packed* scenes with 5 objects placed upright on a table without occlusion, and 2) *pile* scenes where objects were randomly poured into a bin, creating challenging environments at varying occlusion levels with 5, 10, and 15 objects. We used 15 unseen YCB objects [37] as test items, ensuring no overlap with any training sets for fair assessment. For

each attempt, models first predict grasps from the RealSense D435. Then we apply collision detection [7], [18] and background filtering (5cm threshold from segmented foreground objects using [31]). We executed the highest-confidence grasp among the remaining grasps using the Robotiq gripper on a UR5. For 5-object scenes, we standardized arrangements to minimize variance across methods, while for scenes with more than 5 objects, we used identical object sequences with random placement. We conducted 20 rounds for each method and setup, setting maximum consecutive failures at 3 ($\leq$10 objects) and 5 (15 objects) per round.

TABLE IV: **Real-world grasping results.** We report grasp success rate with successful grasps/total trials (in brackets).

Method	Train Data	Packed	Pile		
		5 objects	5 objects	10 objects	15 objects
AnyGrasp	GraspNet-1B++ [9]	62.7(84/134)	66.4(81/122)	60.7(131/216)	59.6(252/423)
Contact-GraspNet	ACRONYM [29]	32.1(34/106)	27.8 (25/90)	34.2 (42/123)	25.7 (46/179)
	GraspNet-1B [7], [18]	77.5(93/120)	62.5(80/128)	58.8(124/211)	54.9(230/419)
	GraspClutter6D	**93.4**(99/106)	**77.2**(95/123)	**74.0**(174/235)	**67.9**(287/423)

Fig. 9: Grasp predictions (top 30) in a **real-world pile scene**.

Real-world Results. Table IV and Fig. 9 present a comparison of grasping performance in the real world. We observed that Contact-GraspNet trained on GraspClutter6D consistently outperformed other methods in all cases. Notably, the network trained on our dataset outperformed AnyGrasp in 15-object scenes (67.9% and 59.6%), highlighting GraspClutter6D as a valuable training source for grasping in complex cluttered environments. However, achieving robust grasping in highly cluttered scenes remains a significant challenge, offering an exciting direction for future research.

B. Impact of GraspClutter6D on Robotic Perception

Setup. We evaluated the effectiveness of GraspClutter6D as a training resource for perception tasks in comparison with GraspNet-1B. We used the standard YCB-Video [4] dataset as our test set (21K images) and trained Mask2Former [2] for segmentation and FFB6D [5] for 6D pose estimation using three different training configurations: (1) YCB-Video only (110K images), (2) YCB-Video + GraspNet-1B (56K additional images), and (3) YCB-Video + GraspClutter6D (14K additional images). We followed the standard evaluation protocols using COCO [50] metrics for segmentation, and ADD metrics [4] for pose estimation.

Results. Tables V and VI present segmentation and pose estimation performance on YCB-Video across different training configurations. Models trained with GraspClutter6D yield substantially greater performance improvements despite using 4$\times$ fewer additional training images (14K vs 56K). These results demonstrate that our dataset provides enhanced training benefits through increased complexity and diversity.

TABLE V: **Segmentation** on YCB-Video (Mask2Former).

Train dataset	AP	AP_{50}	AP_{75}	AR
YCB-Video [4]	57.4	73.1	66.4	60.8
+ GraspNet-1B [7], [18]	61.9 (+4.5)	80.4 (+7.3)	71.0 (+4.6)	66.3 (+5.5)
+ GraspClutter6D	**66.1** (+8.7)	**87.0** (+13.9)	**74.7** (+8.3)	**69.2** (+8.4)

TABLE VI: **Pose estimation** on YCB-Video (FFB6D).

Train dataset	ADD-S	ADD-S
YCB-Video [4]	63.48	81.46
+ GraspNet-1B [7], [18]	64.24 (+0.76)	85.09 (+3.63)
+ GraspClutter6D	**68.80 (+5.32)**	**85.76 (+4.30)**

C. Instance Segmentation Benchmarks

Setup. We evaluated segmentation performance on unknown objects in cluttered scenes using state-of-the-art models: Mask R-CNN [1], Cascade Mask R-CNN [51], and Mask2Former [2]. All models utilized a ResNet-50 FPN [52], [53] backbone with the standard 1$\times$ training schedule and were trained and tested on GraspClutter6D using a cross-object setup, where training and test sets contain different objects. We also evaluated Grounded-SAM [54], a generalist foundation model trained on web-scale datasets, on GraspClutter6D without fine-tuning. For Grounded-SAM, we used a consistent prompt ("an object") and filtered out detections larger than half the image size to reduce false positives. We report average precision and recall using COCO metrics [50].

TABLE VII: **Instance segmentation benchmarks.** $*$ denotes foundation models without domain-specific fine-tuning.

Method	AP	AP_{50}	AP_{75}	AR
Grounded-SAM$*$ [54]	16.2	22.3	17.9	**55.3**
Mask R-CNN [1]	35.2	59.8	35.7	43.3
Cascade R-CNN [51]	35.3	60.4	36.0	44.8
Mask2Former [2]	**43.5**	**69.0**	**44.9**	52.2

Results. Table VII presents segmentation performance on GraspClutter6D. Mask2Former achieved the highest AP of 43.5, outperforming other models. Grounded-SAM demonstrated the highest recall of AR 55.3, but its low precision (AP of 16.2%) reveals limitations in accurately delineating instance boundaries and distinguishing between individual objects in cluttered environments. These results suggest that domain-specific models remain more effective for segmentation in complex environments, though foundation models with adaptation strategies merit further exploration.

D. 6D Object Pose Estimation Benchmarks

Setup. We evaluated object pose estimation on the standard 21 YCB-Video objects [4], [37] in GraspClutter6D using two types of approaches. First, models specialized for known objects—FFB6D [5] and GDR-Net [55]—were trained on 19K images (intra-object setup) with their default configurations. Second, foundation models designed for

novel object generalization—MegaPose [32] and FoundationPose [6]—were evaluated, which were trained on large-scale synthetic datasets (2M and 1.2M images, respectively). We assessed performance using the ADD(-S) and the ADD-S metric [56], [57]. Following established protocols [4], we computed the area under the accuracy-threshold curve up to 10 cm. Results are categorized by occlusion levels based on instance visibility v: low ($v \geq 0.9$), medium ($0.9 > v \geq 0.6$), and high occlusion ($0.6 > v$). For fair comparison, all methods utilized 2D detections from Mask R-CNN [1], except FFB6D which jointly performs segmentation and pose estimation.

TABLE VIII: **Pose estimation results.** ADD(-S)/ADD-S are reported across occlusion levels. $*$: foundation models.

Method	All	Low occ.	Medium occ.	High occ.
FFB6D [5]	36.2 / 59.3	45.1 / 69.5	35.1 / 59.5	13.2 / 26.8
GDR-Net [55]	44.5 / 59.4	52.6 / 69.0	48.2 / 64.0	27.7 / 39.3
MegaPose$*$ [32]	69.6 / 78.6	**77.9** / 86.7	**71.5** / 80.5	42.3 / 51.7
FoundationPose$*$ [6]	**70.5** / **92.3**	76.2 / **94.1**	70.0 / **92.8**	**55.0** / **85.9**

Results. Table VIII presents object pose estimation results on GraspClutter6D. Foundation models substantially outperformed domain-specific methods. These findings suggest that training on large-scale data is effective in improving generalizability for object pose estimation. All methods showed consistent performance degradation with increasing occlusion levels (e.g., MegaPose: 77.9 to 42.3 ADD(-S)), indicating that occlusion handling remains a persistent challenge.

E. 6-DoF Grasp Detection Benchmarks

Setup. We evaluated four methods: Contact-GraspNet [8], GraspNet-Baseline [7], ScaleBalancedGrasp [26], and EconomicGrasp [27]. All models were trained on GraspNet-1B [7], [18] with 100 train scenes and evaluated on (1) GraspNet-1B with 90 test scenes and (2) GraspClutter6D with 235 test scenes, cross-object setup. We compute the standard average precision (AP_μ) [18], which measures the average success rate of the top 50 predicted grasps at friction coefficient μ using the force-closure metric [25], [46]. We report **AP** as a primary metric, averaging AP_μ across friction coefficients from 0.2 to 1.2 at 0.2 intervals. A RealSense D435 camera was used in all datasets. To remove background-induced ambiguity, all models used workspace-cropped inputs and were evaluated for only foreground grasps ($\leq$5cm from target objects).

TABLE IX: **6-DoF grasp detection benchmark results.** (Train: GraspNet-1B, Test: GraspNet-1B, GraspClutter6D)

Method	GraspNet-1B [18]			GraspClutter6D		
	AP	$AP_{0.8}$	$AP_{0.4}$	**AP**	$AP_{0.8}$	$AP_{0.4}$
Contact-GraspNet [8]	19.07	24.34	11.05	10.73	13.18	5.75
GraspNet-Baseline [7]	23.43	27.81	19.56	17.95	21.54	13.53
ScaleBalancedGrasp [26]	44.85	53.31	39.31	**22.69**	**27.57**	**16.95**
EconomicGrasp [27]	**51.63**	**61.55**	**43.72**	21.67	26.62	15.90

Results. Table IX compares the performance of grasping methods across datasets. All methods demonstrate significant performance degradation on GraspClutter6D compared to GraspNet-1B, with EconomicGrasp exhibiting a 29.96 AP reduction. This performance gap indicates that GraspClutter6D can serve as a challenging grasping benchmark, highlighting substantial room for improvement in real-world clutter.

V. CONCLUSION

We presented GraspClutter6D, a dataset for robotic grasping and perception in real-world cluttered environments. Our benchmark evaluations revealed a significant gap, as existing state-of-the-art methods show degraded performance in our high-complexity scenes. We demonstrated that GraspClutter6D serves as an effective training resource; models trained on our dataset achieved substantially improved grasping and perception, highlighting its value in developing robust systems to close this gap. For future work, we will focus on gripper-aware and reactive grasping motions, as we observed that simple approach trajectories often cause failures despite accurate grasp pose detection. By publicly releasing our dataset and toolkits, we hope this dataset will serve as a foundation for advancing robotic manipulation in cluttered environments.

REFERENCES

[1] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proc. IEEE Int. Conf. Comput. Vis*, 2017, pp. 2961–2969.

[2] B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, and R. Girdhar, "Masked-attention mask transformer for universal image segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 1290–1299.

[3] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo *et al.*, "Segment anything," in *Proc. IEEE Int. Conf. Comput. Vis*, 2023, pp. 4015–4026.

[4] Y. Xiang, T. Schmidt, V. Narayanan, and D. Fox, "Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes," *Robotics: Science and Systems XIV*, 2018.

[5] Y. He, H. Huang, H. Fan, Q. Chen, and J. Sun, "Ffb6d: A full flow bidirectional fusion network for 6d pose estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 3003–3013.

[6] B. Wen, W. Yang, J. Kautz, and S. Birchfield, "Foundationpose: Unified 6d pose estimation and tracking of novel objects," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 17 868–17 879.

[7] H.-S. Fang, C. Wang, M. Gou, and C. Lu, "Graspnet-1billion: A large-scale benchmark for general object grasping," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 11 444–11 453.

[8] M. Sundermeyer, A. Mousavian, R. Triebel, and D. Fox, "Contact-graspnet: Efficient 6-dof grasp generation in cluttered scenes," in *Proc. IEEE Int. Conf. Robot. Automat.* IEEE, 2021, pp. 13 438–13 444.

[9] H.-S. Fang, C. Wang, H. Fang, M. Gou, J. Liu, H. Yan, W. Liu, Y. Xie, and C. Lu, "Anygrasp: Robust and efficient grasp perception in spatial and temporal domains," *IEEE Trans. Robot.*, vol. 39, no. 5, pp. 3929–3945, 2023.

[10] C. Rennie, R. Shome, K. E. Bekris, and A. F. De Souza, "A dataset for improved rgbd-based object detection and pose estimation for warehouse pick-and-place," *IEEE Robot. Automat. Lett.*, vol. 1, no. 2, pp. 1179–1185, 2016.

[11] A. Zeng, K.-T. Yu, S. Song, D. Suo, E. Walker, A. Rodriguez, and J. Xiao, "Multi-view self-supervised deep learning for 6d pose estimation in the amazon picking challenge," in *Proc. IEEE Int. Conf. Robot. Automat.* IEEE, 2017, pp. 1386–1383.

[12] T. Hodan, P. Haluza, Š. Obdržálek, J. Matas, M. Lourakis, and X. Zabulis, "T-less: An rgb-d dataset for 6d pose estimation of texture-less objects," in *Proc. Winter Conf. Appl. Comput. Vis.* IEEE, 2017, pp. 880–888.

[13] M. Suchi, T. Patten, D. Fischinger, and M. Vincze, "Easylabel: A semi-automatic pixel-wise object annotation tool for creating robotic rgb-d datasets," in *Proc. IEEE Int. Conf. Robot. Automat.* IEEE, 2019, pp. 6678–6684.

[14] M. Gilles, Y. Chen, E. Z. Zeng, Y. Wu, K. Furmans, A. Wong, and R. Rayyes, "Metagraspnetv2: All-in-one dataset enabling fast and reliable robotic bin picking via object relationship reasoning and dexterous grasping," *IEEE Trans. Autom. Sci. Eng.*, 2023.

[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 *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 5610–5619.

[16] R. Newbury, M. Gu, L. Chumbley, A. Mousavian, C. Eppner, J. Leitner, J. Bohg, A. Morales, T. Asfour, D. Kragic *et al.*, "Deep learning approaches to grasp synthesis: A review," *IEEE Trans. Robot.*, vol. 39, no. 5, pp. 3994–4015, 2023.

[17] D. Bauer, P. Hönig, J.-B. Weibel, J. García-Rodríguez, M. Vincze *et al.*, "Challenges for monocular 6d object pose estimation in robotics," *IEEE Trans. Robot.*, 2024.

[18] H.-S. Fang, M. Gou, C. Wang, and C. Lu, "Robust grasping across diverse sensor qualities: The graspnet-1billion dataset," *Int. J. Robot. Res.*, vol. 42, no. 12, pp. 1094–1103, 2023.

[19] 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 *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 22 498–22 508.

[20] L. Chen, H. Yang, C. Wu, and S. Wu, "Mp6d: An rgb-d dataset for metal parts' 6d pose estimation," *IEEE Robotics and Automation Letters*, vol. 7, no. 3, pp. 5912–5919, 2022.

[21] Y. You, K. Xiong, Z. Yang, Z. Huang, J. Zhou, R. Shi, Z. Fang, A. W. Harley, L. Guibas, and C. Lu, "Pace: A large-scale dataset with pose annotations in cluttered environments," in *Proc. Eur. Conf. Comput. Vis.* Springer, 2024, pp. 473–489.

[22] S. Ainetter and F. Fraundorfer, "End-to-end trainable deep neural network for robotic grasp detection and semantic segmentation from rgb," in *Proc. IEEE Int. Conf. Robot. Automat.* IEEE, 2021, pp. 13 452–13 458.

[23] J. Glover and S. Popovic, "Bingham procrustean alignment for object detection in clutter," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.* IEEE, 2013, pp. 2158–2165.

[24] A. Ten Pas, M. Gualtieri, K. Saenko, and R. Platt, "Grasp pose detection in point clouds," *Int. J. Robot. Res.*, vol. 36, no. 13-14, pp. 1455–1473, 2017.

[25] J. Mahler, J. Liang, S. Niyaz, M. Laskey, R. Doan, X. Liu, J. Aparicio, and K. Goldberg, "Dex-net 2.0: Deep learning to plan robust grasps with synthetic point clouds and analytic grasp metrics," *Robotics: Science and Systems XIII*, 2017.

[26] H. Ma and D. Huang, "Towards scale balanced 6-dof grasp detection in cluttered scenes," in *Proc. Conf. Robot Learn.* PMLR, 2023, pp. 2004–2013.

[27] X.-M. Wu, J.-F. Cai, J.-J. Jiang, D. Zheng, Y.-L. Wei, and W.-S. Zheng, "An economic framework for 6-dof grasp detection," in *Proc. Eur. Conf. Comput. Vis.* Springer, 2024, pp. 357–375.

[28] J. Mahler, M. Matl, V. Satish, M. Danielczuk, B. DeRose, S. McKinley, and K. Goldberg, "Learning ambidextrous robot grasping policies," *Science Robotics*, vol. 4, no. 26, p. eaau4984, 2019.

[29] C. Eppner, A. Mousavian, and D. Fox, "Acronym: A large-scale grasp dataset based on simulation," in *Proc. IEEE Int. Conf. Robot. Automat.* IEEE, 2021, pp. 6222–6227.

[30] C. Xie, Y. Xiang, A. Mousavian, and D. Fox, "Unseen object instance segmentation for robotic environments," *IEEE Trans. Robot.*, vol. 37, no. 5, pp. 1343–1359, 2021.

[31] S. Back, J. Lee, T. Kim, S. Noh, R. Kang, S. Bak, and K. Lee, "Unseen object amodal instance segmentation via hierarchical occlusion modeling," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2022, pp. 5085–5092.

[32] Y. Labbé, L. Manuelli, A. Mousavian, S. Tyree, S. Birchfield, J. Tremblay, J. Carpentier, M. Aubry, D. Fox, and J. Sivic, "Megapose: 6d pose estimation of novel objects via render & compare," in *Proc. Conf. Robot Learn.* PMLR, 2023, pp. 715–725.

[33] A. Richtsfeld, T. Mörwald, J. Prankl, M. Zillich, and M. Vincze, "Segmentation of unknown objects in indoor environments," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.* IEEE, 2012, pp. 4791–4796.

[34] C. Mitash, F. Wang, S. Lu, V. Terhuja, T. Garaas, F. Polido, and M. Nambi, "Armbench: An object-centric benchmark dataset for robotic manipulation," in *Proc. IEEE Int. Conf. Robot. Automat.* IEEE, 2023, pp. 9132–9139.

[35] I. Lenz, H. Lee, and A. Saxena, "Deep learning for detecting robotic grasps," *Int. J. Robot. Res.*, vol. 34, no. 4-5, pp. 705–724, 2015.

[36] A. Depierre, E. Dellandréa, and L. Chen, "Jacquard: A large scale dataset for robotic grasp detection," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.* IEEE, 2018, pp. 3511–3516.

[37] B. Calli, A. Walsman, A. Singh, S. Srinivasa, P. Abbeel, and A. M. Dollar, "Benchmarking in manipulation research: Using the yale-cmu-berkeley object and model set," *IEEE Robot. Auto. Mag.*, vol. 22, no. 3, pp. 36–52, 2015.

[38] S. Tyree, J. Tremblay, T. To, J. Cheng, T. Mosier, J. Smith, and S. Birchfield, "6-dof pose estimation of household objects for robotic manipulation: An accessible dataset and benchmark," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.* IEEE, 2022, pp. 13 081–13 088.

[39] N. Correll, K. E. Bekris, D. Berenson, O. Brock, A. Causo, K. Hauser, K. Okada, A. Rodriguez, J. M. Romano, and P. R. Wurman, "Analysis and observations from the first amazon picking challenge," *IEEE Trans. Autom. Sci. Eng.*, vol. 15, no. 1, pp. 172–188, 2016.

[40] S. Garrido-Jurado, R. Muñoz-Salinas, F. J. Madrid-Cuevas, and M. J. Marín-Jiménez, "Automatic generation and detection of highly reliable fiducial markers under occlusion," *Pattern Recognition*, vol. 47, no. 6, pp. 2280–2292, 2014.

[41] F. Rameau, J. Park, O. Bailo, and I. S. Kweon, "Mc-calib: A generic and robust calibration toolbox for multi-camera systems," *Computer Vision and Image Understanding*, vol. 217, p. 103353, 2022.

[42] V. Lepetit, F. Moreno-Noguer, and P. Fua, "Ep n p: An accurate o (n) solution to the p n p problem," *Int. J. Comput. Vis*, vol. 81, pp. 155–166, 2009.

[43] D. DeTone, T. Malisiewicz, and A. Rabinovich, "Superpoint: Self-supervised interest point detection and description," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2018, pp. 224–236.

[44] P.-E. Sarlin, D. DeTone, T. Malisiewicz, and A. Rabinovich, "Superglue: Learning feature matching with graph neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 4938–4947.

[45] S. Rusinkiewicz and M. Levoy, "Efficient variants of the icp algorithm," in *Proceedings third international conference on 3-D digital imaging and modeling*. IEEE, 2001, pp. 145–152.

[46] V.-D. Nguyen, "Constructing force-closure grasps," *Int. J. Robot. Res.*, vol. 7, no. 3, pp. 3–16, 1988.

[47] S. Hinterstoisser, V. Lepetit, S. Ilic, S. Holzer, G. Bradski, K. Konolige, and N. Navab, "Model based training, detection and pose estimation of texture-less 3d objects in heavily cluttered scenes," in *Proc. Asian Conf. Comput. Vis.* Springer, 2013, pp. 548–562.

[48] E. Coumans and Y. Bai, "Pybullet, a python module for physics simulation for games, robotics and machine learning," 2016.

[49] M. Breyer, J. J. Chung, L. Ott, R. Siegwart, and J. Nieto, "Volumetric grasping network: Real-time 6 dof grasp detection in clutter," in *Proc. Conf. Robot Learn.* PMLR, 2021, pp. 1602–1611.

[50] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.* Springer, 2014, pp. 740–755.

[51] Z. Cai and N. Vasconcelos, "Cascade r-cnn: High quality object detection and instance segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 5, pp. 1483–1498, 2019.

[52] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[53] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2117–2125.

[54] T. Ren, S. Liu, A. Zeng, J. Lin, K. Li, H. Cao, J. Chen, X. Huang, Y. Chen, F. Yan *et al.*, "Grounded sam: Assembling open-world models for diverse visual tasks," *arXiv preprint arXiv:2401.14159*, 2024.

[55] G. Wang, F. Manhardt, F. Tombari, and X. Ji, "Gdr-net: Geometry-guided direct regression network for monocular 6d object pose estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 16 611–16 621.

[56] S. Hinterstoisser, V. Lepetit, S. Ilic, S. Holzer, G. Bradski, K. Konolige, and N. Navab, "Model based training, detection and pose estimation of texture-less 3d objects in heavily cluttered scenes," in *Proc. Asian Conf. Comput. Vis.* Springer, 2012, pp. 548–562.

[57] T. Hodaň, J. Matas, and Š. Obdržálek, "On evaluation of 6d object pose estimation," in *Proc. Eur. Conf. Comput. Vis. Workshops*. Springer, 2016, pp. 606–619.