DexGraspNet 2.0: Learning Generative Dexterous Grasping in Large-scale Synthetic Cluttered Scenes

Jialiang Zhang[1,2,*] Haoran Liu[1,2,*] Danshi Li[2,*] Xinqiang Yu[2,*]
Haoran Geng[1,2] Yufei Ding[1,2] Jiayi Chen[1,2] He Wang[1,2,3,†]

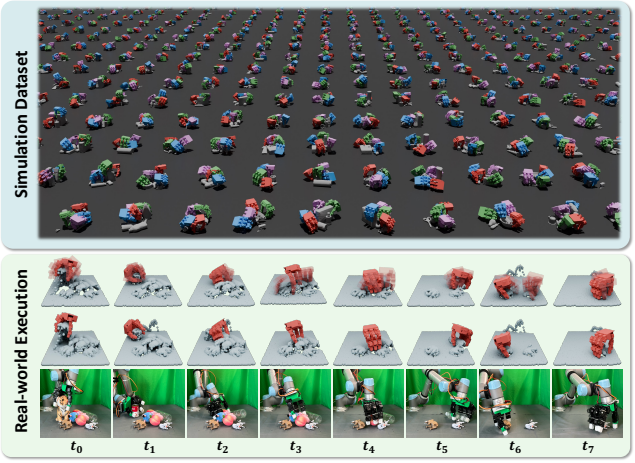

Fig. 1: **Overview.** We propose a comprehensive benchmark named DexGraspNet 2.0 for dexterous grasping in synthetic cluttered scenes. Beyond dataset construction, our proposed method which leverages a generative model conditioned on local features achieves SOTA on DexGraspNet 2.0 and a 90.70% real-world success rate.

***Abstract*—Grasping in cluttered scenes remains highly challenging for dexterous hands due to the scarcity of data. To address this problem, we present a large-scale synthetic dataset, encompassing 1319 objects, 8270 scenes, and 426 million grasps. Beyond benchmarking, we also explore data-efficient learning strategies from grasping data. We reveal that the combination of a generative model conditioned on local features and a grasp dataset that emphasizes complex scene variations is key to achieving effective generalization. Our proposed generative method outperforms all baselines in simulation experiments. Furthermore, it demonstrates zero-shot sim-to-real transfer through test-time depth restoration, attaining 90.70% real-world dexterous grasping success rate, showcasing the robust potential of utilizing fully synthetic training data.**

I. Introduction

Recent years have witnessed significant advancements in dexterous grasping datasets [25, 21, 9] and algorithms [22, 4] for single objects. However, extending these advancements to cluttered scenes poses a formidable challenge due to data scarcity. Existing datasets are either too small [12], contain loosely placed objects [12, 27], or rely on naive search methods [12, 10], hindering algorithm development. Furthermore, this slow progress in dataset research obscures the scale requirements of scenes, grasps, and objects needed for effective generalization.

In this study, we present DexGraspNet 2.0, a large-scale synthetic benchmark for robotic dexterous grasping in cluttered scenes. The dataset comprises 8270 scenes and 426.6 million grasp labels for the LEAP hand [16]. All grasps are synthesized via an optimization process aimed at achieving force closure [2, 25] to ensure diversity and quality.

In addition to data, learning grasping in cluttered scenes presents its own challenges. Firstly, the intricate scene landscapes contribute to a highly complicated distribution of valid grasps, potentially confusing networks. Previous regression methods [10] that directly regress grasp parameters often converge to a mean or median pose in such complex distributions, causing penetration or inaccurate contacts. Secondly, the observation space of cluttered scenes greatly surpasses that for grasping single objects, demanding higher generalization efficiency. Typical grasping approaches for single objects [7, 26] use the global feature vector to predict grasps, necessitating extensive object-level variations to grasp novel objects. Their direct application in cluttered scenarios could significantly impede generalization to new scenes.

To address these challenges, we propose a system that leverages a generative model conditioned on local features to predict grasp pose distribution. Firstly, employing a generative model allows our system to handle the multimodality of grasp distributions more effectively, enhancing output quality. Secondly, by conditioning on local features, our system better exploits the dataset's diverse variations in local geometries, boosting generalization to new objects and scenes.

We acknowledge that our approach is not groundbreaking in its utilization of local features for grasping [24, 10]. However, our contribution lies in offering an intuitive analysis of why such a design can generalize effectively without the need for extensive scene data. Additionally, while the integration of generative models into grasping systems is not novel [7, 26], our work is, to our knowledge, the first to combine this approach with local feature conditioning and validate its effectiveness, paving the way for future advancements.

To comprehensively evaluate our method, we perform simulation experiments on DexGrasNet 2.0, where our model outperforms all baselines. Additionally, we scale down the dataset to identify the turning point for generalization. Finally,

[1]CFCS, School of CS, Peking University
[2]Galbot
[3]Beijing Academy of Artificial Intelligence
[*]Equal contribution
[†]Corresponding author: hewang@pku.edu.cn

with the aid of test-time depth restoration [18], our model achieves a 90.70% success rate in cluttered dexterous grasping in real-world scenarios, confirming the practicality of our system, which trains on fully synthetic data.

II. Related Work

1) Dexterous Grasping Datasets: Creating comprehensive dexterous grasping datasets presents challenges due to the high dimensionality involved. Some studies [23] have utilized teleoperation systems to collect such datasets but face scalability issues. Recent advancements in simulation and synthetic data generation have significantly increased dataset sizes. Sampling-based methods [11] involve sampling grasping poses and selecting optimal ones. Additionally, research [20, 21] suggests using differentiable simulators to generate grasping data. Optimization-based methods [25, 26] refine grasping poses by optimizing a designed energy function. Prior efforts have mainly focused on generating grasping poses for single objects, with limited exploration of cluttered scenes. Although attempts have been made to create datasets for cluttered environments [10], these datasets often lack diversity and quality in challenging settings. In contrast, our work introduces the first comprehensive benchmark for synthetic cluttered scenes, employing optimization-based methods to efficiently generate diverse and high-quality grasping datasets.

2) Data-driven Dexterous Grasping: Data-driven methods utilize these synthetic datasets to learn grasp pose prediction from object point clouds or depth images. Research in this area typically falls into three categories: sampling-based methods, regression-based methods, and generative model approaches. Sampling-based methods [14, 27] often face challenges related to sample efficiency and accuracy. Regression-based methods [4] struggle with handling multimodal data distributions effectively. In contrast, generative models excel in learning data distributions and generating diverse grasping poses when trained on large-scale datasets. Previous works [7, 26] have employed conditional generative models such as CVAE and normalizing flow to learn dexterous grasping, primarily focusing on single-object scenarios. Other studies [27, 12] have explored grasping in cluttered environments; however, they often employ suboptimal designs and feature test scenes where objects are loosely positioned on the table.

III. DexGraspNet 2.0 Benchmark

We present a comprehensive dexterous grasping benchmark, which is a combination of 1319 diverse objects, 8270 cluttered scenes, and 426M grasps in all scenes.

A. Object Collection and Scene Synthesis

For training, we generate 7600 synthetic cluttered scenes using all 60 training objects from [5]. Then we collect 1259 unseen ojbects from [5, 1] and create 670 testing scenes with all 1319 objects. In each scene, 1 to 11 objects are piled within an approximately 30 by 50 cm area, and depth images are rendered from 256 different views. Among the 7600 training scenes, 100 are directly adopted from [5], which are densely packed (containing 8 to 11 objects each). The other 7500 training scenes have a random number of objects. For further details, please refer to our supp.

B. Dexterous Grasp Annotation

We employ a two-stage pipeline to annotate dexterous grasp labels within cluttered training scenes. We first synthesize grasp labels for single objects using our modified implementation of [2, 25] and then leverage the IsaacGym simulator [13] to filter out unstable ones (with friction set to 0.2). Then for each scene, we gather grasps from all objects and retain the collision-free ones. We synthesize approximately 1.9M stable grasps for each object (190M in total), resulting in about 56K collision-free grasps for each scene (426M in total). For more details, please refer to our supp.

C. Dexterous Grasp Evaluation in Simulation

We evaluate various models by their success rates in the IsaacGym [13] simulator. For each test scene, a model receives a single-view depth point cloud and produces a grasp pose. If capable of generating multiple grasps, the model must select the best proposal. A grasp is deemed successful if it can lift an object within the simulator. The friction coefficient is set to 0.2, consistent with the dataset's filtering procedure. We design six test groups consisting of densely, randomly, and loosely packed scenes with objects from GraspNet-1Billion [5], as well as these three types of packed scenes with 1231 objects from ShapeNet [1]. For more details, please consult the supp.

IV. Cluttered Generative Dexterous Grasping

We design a two-stage method to generate dexterous grasp poses in cluttered scenes: (1) a seed point proposal module that identifies graspable regions and extracts point-wise local features, and (2) a grasp pose generation module that models grasp pose distributions conditioned on local features. The combination of the generative model with local conditioning enables our network to learn from numerous local geometry variations in the dataset, which greatly enhances generalization efficiency. We will first introduce the inference process in Sec. IV-A and Sec. IV-B, and then explain the training process in Sec. IV-C. The entire architecture is demonstrated in Fig. 2.

A. Seed Point Proposal

Inspired by [24], the seed point proposal module extracts point-wise features f from a single-view depth point cloud P and identifies graspable regions by generating object segmentation score O and graspness score GS for each point. Based on these, we select a subset of high-scoring points, termed seed points, whose local features are fed into the subsequent grasp generation module, achieving more efficient generalization than conditioning on global features [7, 26].

Ground-truth Graspness Definition. For each training scene, we define the graspness score GS for every point p on the surface of objects, indicating the level of graspability in its surrounding area. In essence, this score is computed by allowing each grasp label to "vote" for nearby points within

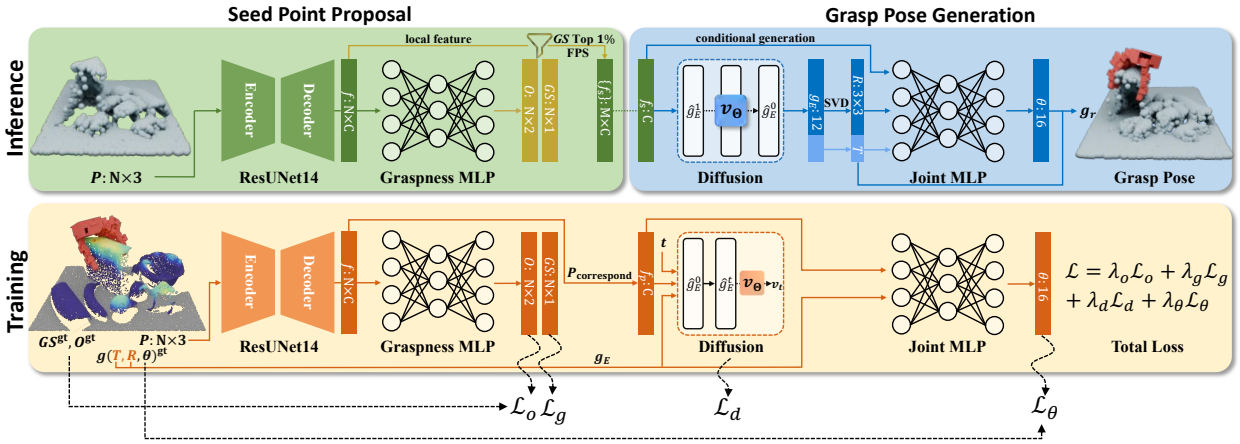

Fig. 2: **Method architecture.** Our method leverages a generative model conditioned on local features and models the distribution of grasp poses (T,R,θ) in a decomposed way. Inference process: The model receives a single-view depth point cloud and generates multiple grasps (only one is visualized). Training process: The model takes the depth point cloud and ground-truth annotations to learn data distribution by optimizing loss terms.

reach of the palm, following a heuristic rule. Subsequently, GS_p is defined as the logarithm of the sum of all voted values. Empirically, GS_p reflects the abundance of valid grasps in its vicinity. And these definition details can be found in the supp.

Graspness Prediction and Seed Point Sampling. To extract local features and predict point-wise graspness, we utilize the same graspness network as described in [24]. Given the scene point cloud $P \in \mathbb{R}^{N\times 3}$, we employ a ResUNet14 built upon MinkowskiEngine to extract a feature vector f_p for each point $p \in \mathbb{R}^3$. This feature is then fed into an MLP to predict p's graspness score GS_p and object segmentation score GS_p, which classifies whether p belongs to an object or the table. Finally, object points that rank top 1% in GS are selected and downsampled to $M = 1024$ points via FPS (farthest point sampling). These points, denoted as $S = \{s\}$, are termed seed points. Their feature vectors $\{f_s\}$ and graspness scores $\{GS_s\}$ are used for subsequent grasp pose generation.

B. Grasp Pose Generation

The grasp pose generation module takes a seed point's feature vector f_s, and generates diverse dexterous grasp poses relative to that seed point using a generative model. These grasp poses are then ranked based on their estimated log-likelihoods and the graspness GS_s of the seed point.

Notations and Assumptions. A dexterous grasp pose relative to a seed point is denoted as $g^r = (T, R, \theta)$, where $T \in \mathbb{R}^3$ and $R \in \mathrm{SO}(3)$ represent the wrist pose relative to the seed point, and $\theta \in \mathbb{R}^{\mathrm{DoF}}$ signifies joint angles of the hand ($\mathrm{DoF} = 16$ for LEAP hand [17]). Given a seed point s from a scene point cloud P, all valid grasps near s form a conditional probability distribution $p(T, R, \theta|f_s)$, where f_s is the predicted visual feature of point s. We assume that the distribution of (T, R) conditioned on f_s is multi-moded and complicated, while the distribution of θ conditioned on f_s and (T, R) is single-moded. Therefore, we use a conditional generative model to predict the conditional distribution $p(T, R|f_s)$, and a deterministic model to predict θ from f_s and (T, R).

Predicting Conditional Pose Distribution via Diffusion. We adopt the denoising diffusion probabilistic model [6], a powerful class of probabilistic models widely used in the Euclidean space, to approximate $p(T, R|f_s)$. To embed $p(T, R|f_s)$ into the Euclidean space, we flatten the rotation matrix R and concatenate it with the translation T to get the 12D vector representation g_E of a grasp's wrist pose. Then we learn a conditional denoising model v_Θ to denoise a random 12D Gaussian noise vector $\hat{g}_E^1$ into a valid wrist pose vector $g_E = \hat{g}_E^0$ through an iterative process. Specifically, at each diffusion timestep $t \in [0, 1]$, we feed t, feature vector f_s, and the current noisy sample $\hat{g}_E^t$ to an MLP v_Θ, which then predicts the velocity [15] of the diffusion process. Using the predicted velocity, we denoise $\hat{g}_E^t$ into $\hat{g}_E^{t-\mathrm{d}t}$ by solving an ODE illustrated in [19]. After the last step, the denoised $g_E = \hat{g}_E^0 \in \mathbb{R}^{12}$ is projected back to $\mathrm{SE}(3)$ by applying SVD [8] to the rotation channels. Moreover, we estimate the sample's probability $p(g_E|f_s)$ by solving a PDE introduced in [3, 19], and then empirically rank the sample with a linear combination of $\log p(g_E|f_s)$ and GS_s.

Finger Joint Angle Prediction. After sampling a wrist pose (T, R) from the diffusion model, we input f_s and (T, R) into an MLP to predict the finger joint angles θ, together forming $g^r = (T, R, \theta)$, a generated dexterous grasp pose relative to seed point s. Additionally, our method seamlessly extends to parallel grippers by substituting the joint angles $\theta \in \mathbb{R}^{16}$ with one parameter $w \in \mathbb{R}$ indicating gripper width.

C. Joint Training and Loss Functions

The seed point proposal module and grasp pose generation module are trained jointly. At each gradient step, we randomly sample $D = 8$ scenes from our dataset and select a rendering view for each scene. The depth point cloud of a scene is denoted as $P \in \mathbb{R}^{N\times 3}$, its corresponding object point mask is $\{O_p^{\mathrm{gt}}\}_N$, and the ground truth point-wise graspness scores are

$\{GS_p^{\text{gt}}\}_N$. All point coordinates are represented in the camera frame. We then sample $B = 64$ random grasp labels $\{g\}_B$ in this scene. For each grasp g, we compute its "corresponding point" p following Sec. IV-A and represent g as a relative grasp pose $(T^{\text{gt}}, R^{\text{gt}}, \theta^{\text{gt}})$ in the reference frame of p.

First, point cloud P is fed into the seed point proposal module to obtain local features $\{f_p\}_N$, object segmentation scores $\{O_{p,0/1}\}_N$, and point-wise graspness scores $\{GS_p\}_N$, after which the object segmentation loss $\mathcal{L}_o$ (Cross-Entropy) and graspness loss $\mathcal{L}_g$ (SmoothL1) are computed.

Next, for each grasp g_j, we collect the local feature f_p of its corresponding point p and pass these to the grasp generation module. The 12D Euclidean representation of the wrist pose g_E undergoes the diffusion process to obtain a noisy sample $\hat{g}_E^t = \sqrt{\overline{\alpha}_t} g_E + \sqrt{1-\overline{\alpha}_t}\epsilon$ and diffusion velocity $v_t = \sqrt{\overline{\alpha}_t}\epsilon - \sqrt{1-\overline{\alpha}_t} g_E$ at some random time step t. The denoising model v_Θ then predicts this velocity, under the supervision of an MSE loss $\mathcal{L}_d$. The joint angle prediction MLP takes feature f_p and the wrist pose $(T^{\text{gt}}, R^{\text{gt}})$, predicts θ, and is supervised by a joint angle loss $\mathcal{L}_\theta$ (Smooth L1).

The total loss is a linear combination of all loss terms: $\mathcal{L} = \lambda_o\mathcal{L}_o + \lambda_g\mathcal{L}_g + \lambda_d\mathcal{L}_d + \lambda_\theta\mathcal{L}_\theta$. The model is trained on one NVIDIA 3090 for 50k iterations, taking about 2 hours.

V. Experiment

A. Baseline comparisons.

We compare our method with HGC-Net [10] and modified versions of GraspTTA [7] and ISAGrasp [4] on DexGraspNet 2.0, with simulation success rates shown in Tab. I. Our method leverages a diffusion model conditioned on local features, which boosts prediction accuracy as well as generalizability, significantly outperforming all baselines.

B. Scaling the Dataset.

We scale down the training data in two ways: (1) by reducing the number of grasps in each scene, and (2) by decreasing the number of training scenes.

Grasps. Our model demonstrates a significant performance increase from 40K to 4M grasps, indicating strong scaling properties. In contrast, the modified ISAGrasp, following the regression approach, only increases at a 10% success rate, saturating at 66.1%, when the number of grasps rises from 40K to 400M. This suggests that regressive methods may struggle to effectively leverage increased grasp data due to the complexity of the data distribution.

Scenes. Our model continuously improves in performance with the addition of more training scenes, indicating that it benefits from a larger and more diverse scene dataset. Interestingly, although the total number of training objects never exceeds 60, the model achieves an impressive 85.8% success rate on test scenes containing 1231 novel objects, which implies scenes matter more than objects for our model.

C. Real-World Experiments

We use the collection of 32 objects in our real-world experiments. And, we use a LEAP hand mounted on the UR-5 robot arm for the dexterous grasp experiment and a Franka Panda arm for the gripper experiment. Both of them use an Intel RealSense D435 camera and use [18] for depth restoration. In both settings, we grasp objects in a cluttered scene one by one until the table is cleared. As shown in Tab. II, our method achieves 90.70% and 92.45% success rates for dexterous hand and gripper respectively, outperforming baseline methods.

Method	GraspNet-1Billion		ShapeNet	
	Dense	Loose	Dense	Loose
HGC-Net [10]	46.0	26.7	46.4	30.4
GraspTTA†[7]	62.5	42.8	56.6	46.4
ISAGrasp† [4]	63.4	51.4	64.0	52.7
Ours	**90.6**	**73.2**	**81.0**	**74.2**

TABLE I: **Benchmark for dexterous grasping.** Modified baseline methods are indicated with †. Each **Dense** scene contains 8-11 objects; each **Loose** scene contains 1-2 objects.

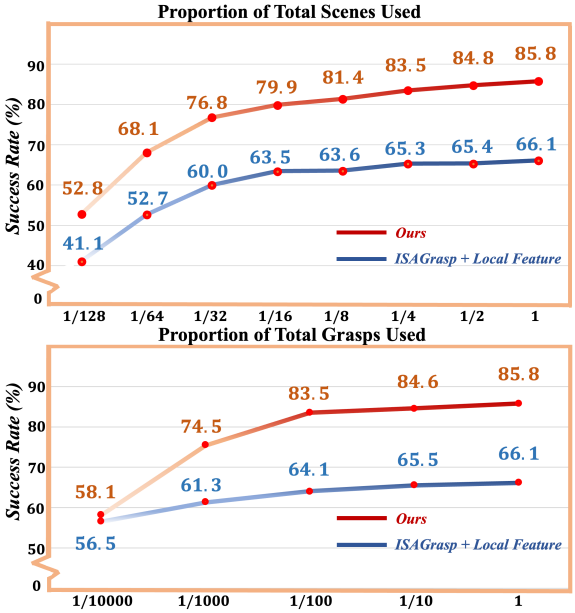

Fig. 3: Scaling the number of grasps/scenes.

	Dexterous Grasp		Gripper Grasp	
Method	HGC-Net	Ours	ASGrasp	Ours
SR (%)	16.44	**90.70**	84.48	**92.45**

TABLE II: Comparison of real-robot success rates.

VI. Conclusion

We present DexGraspNet 2.0, a large-scale benchmark for dexterous grasping in cluttered scenes. Our proposed method outperforms all baselines in simulation and achieves a 90.70% success rate in real-world tests. By scaling our dataset, we identify the turning point for generalization. We also discover that scene complexity is more important for generalization than the number of training objects.

References

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

[2] Jiayi Chen, Yuxing Chen, Jialiang Zhang, and He Wang. Task-oriented dexterous grasp synthesis via differentiable grasp wrench boundary estimator. arXiv preprint arXiv:2309.13586, 2023.

[3] Ricky TQ Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary differential equations. Advances in neural information processing systems, 31, 2018.

[4] Zoey Qiuyu Chen, Karl Van Wyk, Yu-Wei Chao, Wei Yang, Arsalan Mousavian, Abhishek Gupta, and Dieter Fox. Learning robust real-world dexterous grasping policies via implicit shape augmentation. arXiv preprint arXiv:2210.13638, 2022.

[5] Hao-Shu Fang, Chenxi Wang, Minghao Gou, and Cewu Lu. Graspnet-1billion: A large-scale benchmark for general object grasping. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 11444–11453, 2020.

[6] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. Advances in neural information processing systems, 33:6840–6851, 2020.

[7] Hanwen Jiang, Shaowei Liu, Jiashun Wang, and Xiaolong Wang. Hand-object contact consistency reasoning for human grasps generation. In Proceedings of the IEEE/CVF international conference on computer vision, pages 11107–11116, 2021.

[8] Jake Levinson, Carlos Esteves, Kefan Chen, Noah Snavely, Angjoo Kanazawa, Afshin Rostamizadeh, and Ameesh Makadia. An analysis of svd for deep rotation estimation. arXiv preprint arXiv:2006.14616, 2020.

[9] Puhao Li, Tengyu Liu, Yuyang Li, Yiran Geng, Yixin Zhu, Yaodong Yang, and Siyuan Huang. Gendexgrasp: Generalizable dexterous grasping. In 2023 IEEE International Conference on Robotics and Automation (ICRA), pages 8068–8074. IEEE, 2023.

[10] Yiming Li, Wei Wei, Daheng Li, Peng Wang, Wanyi Li, and Jun Zhong. Hgc-net: Deep anthropomorphic hand grasping in clutter. In 2022 International Conference on Robotics and Automation (ICRA), pages 714–720. IEEE, 2022.

[11] Jens Lundell, Enric Corona, Tran Nguyen Le, Francesco Verdoja, Philippe Weinzaepfel, Grégory Rogez, Francesc Moreno-Noguer, and Ville Kyrki. Multi-fingan: Generative coarse-to-fine sampling of multi-finger grasps. In 2021 IEEE International Conference on Robotics and Automation (ICRA), pages 4495–4501. IEEE, 2021.

[12] Jens Lundell, Francesco Verdoja, and Ville Kyrki. Ddgc: Generative deep dexterous grasping in clutter. IEEE Robotics and Automation Letters, 6(4):6899–6906, 2021.

[13] Viktor Makoviychuk, Lukasz Wawrzyniak, Yunrong Guo, Michelle Lu, Kier Storey, Miles Macklin, David Hoeller, Nikita Rudin, Arthur Allshire, Ankur Handa, et al. Isaac gym: High performance gpu-based physics simulation for robot learning. arXiv preprint arXiv:2108.10470, 2021.

[14] Andrew T Miller and Peter K Allen. Graspit! a versatile simulator for robotic grasping. IEEE Robotics & Automation Magazine, 11(4):110–122, 2004.

[15] Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. arXiv preprint arXiv:2202.00512, 2022.

[16] Kenneth Shaw, Ananye Agarwal, and Deepak Pathak. Leap hand: Low-cost, efficient, and anthropomorphic hand for robot learning. arXiv preprint arXiv:2309.06440, 2023.

[17] Kenneth Shaw, Ananye Agarwal, and Deepak Pathak. Leap hand: Low-cost, efficient, and anthropomorphic hand for robot learning. Robotics: Science and Systems (RSS), 2023.

[18] Jun Shi, Yixiang Jin, Dingzhe Li, Haoyu Niu, Zhezhu Jin, He Wang, et al. Asgrasp: Generalizable transparent object reconstruction and grasping from rgb-d active stereo camera. arXiv preprint arXiv:2405.05648, 2024.

[19] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. arXiv preprint arXiv:2011.13456, 2020.

[20] Dylan Turpin, Liquan Wang, Eric Heiden, Yun-Chun Chen, Miles Macklin, Stavros Tsogkas, Sven Dickinson, and Animesh Garg. Grasp'd: Differentiable contact-rich grasp synthesis for multi-fingered hands, 2022.

[21] Dylan Turpin, Tao Zhong, Shutong Zhang, Guanglei Zhu, Jingzhou Liu, Ritvik Singh, Eric Heiden, Miles Macklin, Stavros Tsogkas, Sven Dickinson, and Animesh Garg. Fast-grasp'd: Dexterous multi-finger grasp generation through differentiable simulation, 2023.

[22] Weikang Wan, Haoran Geng, Yun Liu, Zikang Shan, Yaodong Yang, Li Yi, and He Wang. Unidexgrasp++: Improving dexterous grasping policy learning via geometry-aware curriculum and iterative generalist-specialist learning. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 3891–3902, 2023.

[23] Chen Wang, Haochen Shi, Weizhuo Wang, Ruohan Zhang, Li Fei-Fei, and C. Karen Liu. Dexcap: Scalable and portable mocap data collection system for dexterous manipulation, 2024.

[24] Chenxi Wang, Hao-Shu Fang, Minghao Gou, Hongjie Fang, Jin Gao, and Cewu Lu. Graspness discovery in clutters for fast and accurate grasp detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 15964–15973, 2021.

[25] Ruicheng Wang, Jialiang Zhang, Jiayi Chen, Yinzhen Xu, Puhao Li, Tengyu Liu, and He Wang. Dexgraspnet: A large-scale robotic dexterous grasp dataset for general

objects based on simulation. In 2023 IEEE International Conference on Robotics and Automation (ICRA), pages 11359–11366. IEEE, 2023.

[26] Yinzhen Xu, Weikang Wan, Jialiang Zhang, Haoran Liu, Zikang Shan, Hao Shen, Ruicheng Wang, Haoran Geng, Yijia Weng, Jiayi Chen, et al. Unidexgrasp: Universal robotic dexterous grasping via learning diverse proposal generation and goal-conditioned policy. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 4737–4746, 2023.

[27] Lei Zhang, Kaixin Bai, Guowen Huang, Zhaopeng Chen, and Jianwei Zhang. Multi-fingered robotic hand grasping in cluttered environments through hand-object contact semantic mapping. arXiv preprint arXiv:2404.08844, 2024.