GRASP DIFFUSION NETWORK: Learning Grasp Generators from Partial Point Clouds with Diffusion Models in SO(3)×$\mathbb{R}^3$

Joao Carvalho[1], An T. Le[1], Philipp Jahr[1], Qiao Sun[1], Julen Urain[3], Dorothea Koert[1,2], and Jan Peters[1,3,4]

***Abstract*— Grasping objects successfully from a single-view camera is crucial in many manipulation tasks. An approach to solve this problem is to leverage simulation to create large datasets of pairs of objects and grasp poses, and then learn a conditional generative model that can be prompted quickly during deployment. However, the grasp pose data is highly multimodal since there are several ways to grasp an object. In this work, we learn a grasp generative model with diffusion models to sample candidate grasp poses given a partial point cloud of an object. We show in real-world experiments that our approach can grasp several objects from raw depth images with 90% success rate and benchmark it against several baselines. https://sites.google.com/view/graspdiffusionnetwork**

I. INTRODUCTION

We consider the problem of learning a distribution of candidate successful grasps poses of a parallel gripper, given a *partial view* of an object from a depth camera (a partial point cloud). Full point clouds are often impractical to obtain in the real world, especially for grasping scenarios (e.g., if an object is placed on top of a table, we cannot obtain the point cloud of its bottom). The grasp pose data is often not unimodal since several grasp poses can successfully grasp an object. Diffusion models are a good candidate for modeling these distributions because, contrary to other generative models, such as VAEs and GANs [1], [2], they avoid mode collapse, can handle large amounts of data, and are empirically more stable to train [3]. Moreover, they are prolific in many robotic manipulation fields [4]–[8].

Denoising Diffusion Probabilistic Models (DDPM) [9] were designed for Euclidean spaces. However, grasp poses belong to the space of homogeneous transformations SE(3). To adapt diffusion to this space, we decouple the Lie group SE(3) into SO(3)×$\mathbb{R}^3$ and learn a diffusion model in the Lie algebra. As previously shown in [10], [11], properly modeling the diffusion of rotations has several benefits.

Our contributions include: (1) We present the Grasp Diffusion Network (GDN) – a new grasp generative model conditioned on partial point clouds that uses diffusion in the SO(3)×$\mathbb{R}^3$ manifold. (2) We evaluate GDN in the real world using partial point clouds obtained with a RGB-D camera and show that it achieves higher success rates than the baselines.

Corresponding author: Joao Carvalho, joao@robot-learning.de

This work was funded by the German Federal Ministry of Education and Research projects IKIDA (01IS20045) and Software Campus project ROBOSTRUCT (01S23067), and by the German Research Foundation project METRIC4IMITATION (PE 2315/11-1). [1]Computer Science Department, TU of Darmstadt, Germany; [2]Centre for Cognitive Science, TU Darmstadt, Germany; [3]German Research Center for AI (DFKI), Research Department: SAIROL, Darmstadt, Germany; [4]Hessian.AI, Darmstadt, Germany

II. RELATED WORK

For complete surveys on deep learning models for grasping, we refer the readers to [12], [13]. A seminal work in this field is 6-dof GraspNet [14], which uses a conditional variational autoencoder (CVAE) [15] to learn a grasp distribution given partial point clouds from single objects observations. Several works build on top of this model to extend grasping to full scene point clouds [16], [17]. In recent works, score-based and diffusion models have been proposed [7], [18], [19]. SE(3)-DiffusionFields [7] learns a grasp distribution of a parallel gripper with an energy-based model (EBM) via denoising score matching, and uses Langevin dynamics to sample from the learned model, which involves backpropagating through the entire EBM network to obtain a log-probability gradient. Instead, we model the generative model using DDPM, and our network directly computes the gradient as the denoising function output. GraspLDM [19] learns a grasp generative model using latent diffusion in Euclidean space. However, a CVAE is first used to learn the latent space representation.

III. GRASP DIFFUSION NETWORK

Our goal is to approximate $p(\mathcal{G}|\mathcal{C})$, a distribution of a parallel gripper grasp poses $\mathcal{G}$ conditioned on a partial point cloud observation $\mathcal{C}$. A grasp pose $\boldsymbol{G} = (\boldsymbol{t}, \boldsymbol{R})$ is an element of SE(3), where $\boldsymbol{t} \in \mathbb{R}^3$ is a translation and $\boldsymbol{R} \in$ SO(3) a rotation. As DDPM models were derived for the Euclidean space, for the translation part, we use DDPM in $\mathbb{R}^3$, while for the rotation part, we define a diffusion model in SO(3) as in [10]. The denoising posterior is a factorized distribution

$$\begin{aligned} p_{\boldsymbol{\theta}}(\boldsymbol{G}_{i-1}|\boldsymbol{G}_i, i, \boldsymbol{c}) &= \mathcal{IG}_{\text{SO(3)}\times\mathbb{R}^3}(\boldsymbol{G}_{i-1}; \boldsymbol{\mu}_{\boldsymbol{\theta}}(\boldsymbol{G}_i, i, \boldsymbol{c}), \boldsymbol{\Sigma}_i) \\ &= \mathcal{N}\left(\boldsymbol{t}_{i-1}; \boldsymbol{\mu}^{\boldsymbol{t}}_{\boldsymbol{\theta}}, \boldsymbol{\Sigma}_i\right) \mathcal{IG}_{\text{SO(3)}}(\boldsymbol{R}_{i-1}; \boldsymbol{\mu}^{\boldsymbol{R}}_{\boldsymbol{\theta}}, \boldsymbol{\Sigma}_i), \end{aligned}$$

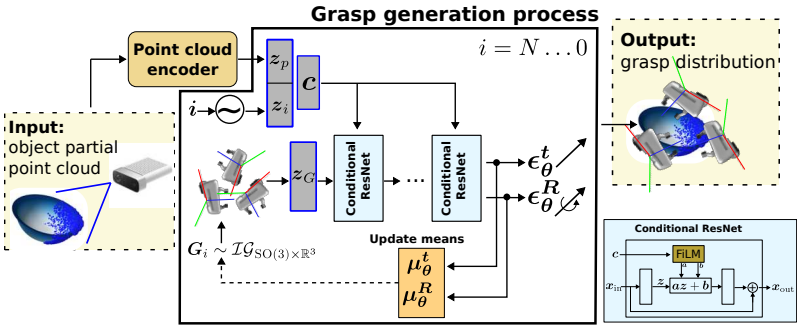
where $\boldsymbol{c}$ is a conditioning variable representing a partial point cloud embedding, and $\mathcal{IG}_{\text{SO(3)}}$ is the isotropic Gaussian distribution in SO(3) [11]. To sample grasp poses, we start from random noise and iteratively sample from this posterior distribution for N steps. For faster inference, we also use Denoising Diffusion Implicit Models (DDIM) [20].

Similar to [9], we learn the joint denoising vector $\boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\boldsymbol{G}_i, i, \boldsymbol{c}) = [\boldsymbol{\epsilon}^{\boldsymbol{t}}_{\boldsymbol{\theta}}, \boldsymbol{\epsilon}^{\boldsymbol{R}}_{\boldsymbol{\theta}}] \in \mathbb{R}^6$, and compute the means with

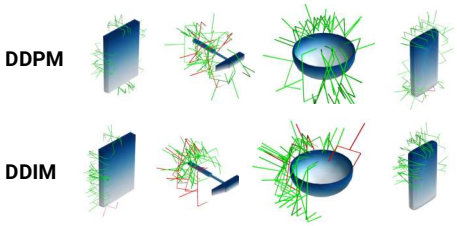
$$\boldsymbol{\mu}^{\boldsymbol{t}}_{\boldsymbol{\theta}} = \frac{1}{\sqrt{\alpha_i}} \left(\boldsymbol{t}_i - \frac{1-\alpha_i}{\sqrt{1-\bar{\alpha}_i}} \boldsymbol{\epsilon}^{\boldsymbol{t}}_{\boldsymbol{\theta}} \right) \tag{1}$$

$$\boldsymbol{\mu}^{\boldsymbol{R}}_{\boldsymbol{\theta}} = \text{Exp}\left(\frac{1}{\sqrt{\alpha_i}} \left(\text{Log}\, \boldsymbol{R}_i - \frac{1-\alpha_i}{\sqrt{1-\bar{\alpha}_i}} \boldsymbol{\epsilon}^{\boldsymbol{R}}_{\boldsymbol{\theta}} \right) \right), \tag{2}$$

for an appropriate diffusion noise schedule α_i. The denoising vector $\boldsymbol{\epsilon}^{\boldsymbol{R}}_{\boldsymbol{\theta}} \in \mathbb{R}^3$ is represented in the Lie algebra of SO(3).



(a) Grasp Diffusion Network inference pipeline



(b) Samples from GDN with DDPM and DDIM

Fig. 1: (a) The input to GDN is a partial point cloud of the object (blue dots), and the output is a distribution of gripper poses by denoising in the SO(3)×$\mathbb{R}^3$ manifold. The denoising posterior means are updated with eqs. (1) and (2). (b) Grasp samples generated with GDN using DDPM and DDIM sampling methods. The colors green and red indicate if the grasp was successful or unsuccessful in simulation.

The Exp and Log operators map elements from the Lie algebra to the Lie group and vice-versa, respectively. The parameters of $\boldsymbol{\epsilon_\theta}(\boldsymbol{G}_i, i, \boldsymbol{c})$ are learned by minimizing the loss

$$\mathcal{L}(\boldsymbol{\theta}) = \mathbb{E}_{i,\boldsymbol{\epsilon},\boldsymbol{c},\boldsymbol{G}_0}\left[\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon_\theta}(\boldsymbol{G}_i, i, \boldsymbol{c})\|_2^2\right], \tag{3}$$

with $i \sim \mathcal{U}(1, N)$, $(\boldsymbol{\epsilon^R}, \boldsymbol{\epsilon^t}) \sim \mathcal{IG}_{\text{SO(3)}\times\mathbb{R}^3}(\boldsymbol{0}, \boldsymbol{I})$, $\boldsymbol{c}$ an object partial point cloud, and $\boldsymbol{G}_0 = (\boldsymbol{t}_0, \boldsymbol{R}_0)$ a grasp sample from the dataset. The noisy sample $\boldsymbol{G}_i = (\boldsymbol{t}_i, \boldsymbol{R}_i)$ is constructed with $\boldsymbol{t}_i = \sqrt{\bar{\alpha}_i}\boldsymbol{t}_0 + (1-\bar{\alpha}_i)\boldsymbol{I}\boldsymbol{\epsilon^t}$, $\boldsymbol{R}_i = \boldsymbol{R}\lambda(\sqrt{\bar{\alpha}_i}, \boldsymbol{R}_0)$, $\boldsymbol{R} \sim \mathcal{IG}_{\text{SO(3)}}(\boldsymbol{I}, \sqrt{1-\bar{\alpha}_i})$, where λ is the geodesic interpolation from the identity rotation $\lambda(\gamma, \boldsymbol{R}) = \exp(\gamma \log(\boldsymbol{R}))$.

Figure 1a shows an overview of our method – GDN. Figure 1b displays grasp samples generated in simulation with slow (DDPM) and faster (DDIM) sampling methods. As expected, DDIM produces less diverse samples. See [21] for more details.

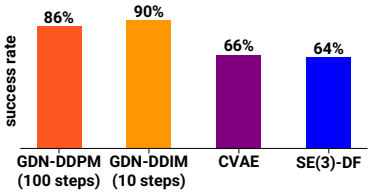
IV. Experiments

To train GDN we use the ACRONYM dataset [22], fully generated in simulation, and choose a subset of everyday objects from 10 categories totaling 567 objects and ≈1M grasps. We use the train/test splits from [16]. The DDPM model uses 100 diffusion steps and a noise cosine schedule. The loss function is optimized with mini-batch gradient descent using the Adam optimizer [23] and a learning rate 3×10^{-4}. In each mini-batch, with a virtual camera we render partial point clouds (with 1024 points) of 32 objects and use 32 grasps per object (a total of 1024 grasps per batch).

To evaluate our models we use – GDN using DDPM, and GDN using DDIM with 10 sampling steps –, and two baselines: a CVAE model similar to [14], [19], and SE(3)-DiffusionFields (SE(3)-DF) [7], which establishes the state-of-the-art for grasp generation with score-models in SE(3).

In this experiment, a robot needs to grasp an object from a table and drop it in a box, using only a partial point cloud view from an external RGB-D camera (Azure Kinect). We bought 10 objects from a local store from the training set categories, displayed in fig. 2a. To calibrate the external camera we used MoveIt [24], and to segment and retrieve the object's partial point cloud, we use the segmentation method FastSAM [25]. We sample 100 candidate grasp poses, remove those colliding with the partial point cloud and the table, and select one randomly. We use a path planner to move the robot

(a)



(b)

Fig. 2: (a) Objects used in the real-world experiments. Each object is placed in 5 different configurations. (b) Average success rates obtained by GDN methods and baselines.

Fig. 3: The figures show overlays of the successful grasp execution sampled with GDN for different real-world objects.

to a pre-grasp and grasp joint goal configurations (computed with inverse kinematics) while avoiding collisions. At the grasp pose, the gripper is closed, and the end-effector moves first 50cm vertically and then to a disposal area, where the object is dropped into a box. We consider a successful grasp if the gripper can hold the object until reaching the box.

Figure 3 shows examples of successful grasp executions sampled with GDN for different objects. In fig. 2b we report that GDN with DDIM sampling obtained an average 90% success rate, which is considerably higher than the baselines CVAE and SE(3)-DF (these results mirror the insights obtained from simulation results). One possible reason to justify the DDPM and DDIM results is that the latter is a deterministic sampling method known to produce more precise samples, albeit less diverse (see fig. 1b).

V. Conclusion

We proposed the Grasp Diffusion Network (GDN), a novel grasp generative model to sample grasp poses given partial point clouds of single objects. GDN encodes the grasp distribution with diffusion in the SO(3)×$\mathbb{R}^3$ manifold. In a grasping scenario with real-world objects, we show that GDN can transfer from simulation to the real world and obtain better success rates than the baselines. In future work, we will expand our method to grasp objects in cluttered scenes.

References

[1] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," in *ICLR*, 2014.

[2] I. Goodfellow, J. Pouget-Abadie, *et al.*, "Generative adversarial nets," in *NeurIPS*, vol. 27, 2014.

[3] R. Bayat, "A study on sample diversity in generative models: Gans vs. diffusion models," in *The First Tiny Papers Track at ICLR 2023, Tiny Papers @ ICLR 2023, Kigali, Rwanda, May 5, 2023*, 2023.

[4] J. Urain, A. Mandlekar, *et al.*, "Deep generative models in robotics: A survey on learning from multimodal demonstrations," 2024.

[5] C. Chi, S. Feng, *et al.*, "Diffusion policy: Visuomotor policy learning via action diffusion," in *R:SS XIX*, 2023.

[6] M. Reuss, M. Li, *et al.*, "Goal-conditioned imitation learning using score-based diffusion policies," in *R:SS XIX*, 2023.

[7] J. Urain, N. Funk, *et al.*, "Se(3)-diffusionfields: Learning smooth cost functions for joint grasp and motion optimization through diffusion," in *IEEE ICRA*, 2023.

[8] J. Carvalho, A. T. Le, *et al.*, "Motion planning diffusion: Learning and planning of robot motions with diffusion models," in *IROS*, 2023.

[9] J. Ho, A. Jain, *et al.*, "Denoising diffusion probabilistic models," in *NeurIPS*, 2020.

[10] A. Leach, S. M. Schmon, *et al.*, "Denoising diffusion probabilistic models on SO(3) for rotational alignment," in *ICLR 2022 Workshop on Geometrical and Topological Representation Learning*, 2022.

[11] Y. Jagvaral, F. Lanusse, *et al.*, "Unified framework for diffusion generative models in SO(3): applications in computer vision and astrophysics," in *AAAI*, 2024.

[12] R. Newbury, M. Gu, *et al.*, "Deep learning approaches to grasp synthesis: A review," *IEEE Trans. Robotics*, vol. 39, no. 5, 2023.

[13] R. Platt, "Grasp learning: Models, methods, and performance," *Annu. Rev. Control. Robotics Auton. Syst.*, vol. 6, 2023.

[14] A. Mousavian, C. Eppner, *et al.*, "6-dof graspnet: Variational grasp generation for object manipulation," in *ICCV*, 2019.

[15] K. Sohn, H. Lee, *et al.*, "Learning structured output representation using deep conditional generative models," in *NeurIPS*, 2015.

[16] M. Sundermeyer, A. Mousavian, *et al.*, "Contact-graspnet: Efficient 6-dof grasp generation in cluttered scenes," in *ICRA*, 2021.

[17] A. Murali, A. Mousavian, *et al.*, "6-dof grasping for target-driven object manipulation in clutter," in *ICRA*, 2020.

[18] P. Song, P. Li, *et al.*, "Implicit grasp diffusion: Bridging the gap between dense prediction and sampling-based grasping," in *8th Annual Conference on Robot Learning*, 2024.

[19] K. R. Barad, A. Orsula, *et al.*, "Graspldm: Generative 6-dof grasp synthesis using latent diffusion models," *CoRR*, vol. abs/2312.11243, 2023.

[20] J. Song, C. Meng, *et al.*, "Denoising diffusion implicit models," in *ICLR*, 2021.

[21] J. Carvalho, A. T. Le, *et al.*, "Grasp diffusion network: Learning grasp generators from partial point clouds with diffusion models in so(3)xr3," 2024.

[22] C. Eppner, A. Mousavian, *et al.*, "ACRONYM: A large-scale grasp dataset based on simulation," in *ICRA*, 2021.

[23] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *ICLR*, 2015.

[24] D. Coleman, I. A. Sucan, *et al.*, "Reducing the barrier to entry of complex robotic software: a moveit! case study," *CoRR*, vol. abs/1404.3785, 2014.

[25] X. Zhao, W. Ding, *et al.*, "Fast segment anything," *CoRR*, vol. abs/2306.12156, 2023.