Leveraging Equivariant Representations of 3D Point Clouds for SO(3)-Equivariant 6-DoF Grasp Pose Generation

Byeongdo Lim*[1], Jongmin Kim*[1], Jihwan Kim[1], Yonghyeon Lee[2], and Frank C. Park[1,3]

***Abstract*— Achieving equivariance in robot learning tasks, particularly in the generation of grasp poses for various objects, has garnered significant attention due to its advantages such as data efficiency, generalization, and robustness. In this paper, we propose GraspECMF (Equivariant Conditional Manifold Flows for grasping), a novel method for SO(3)-equivariant grasp pose generation. Our method leverages SO(3)-equivariant representations of objects to learn the invariant distribution of grasp poses conditioned on the objects. Experimental validation demonstrates that our method outperforms existing methods, showcasing enhanced accuracy in grasp pose distribution learning and resulting in a higher grasp success rate.**

I. INTRODUCTION

In robot learning tasks, the principle of equivariance – that when a symmetry transformation is applied to the environment, the robot's action (e.g., end-effector's pose or trajectory) should undergo a transformation identical to that applied to the environment – has drawn significant attention for its advantages in enhancing data efficiency, generalization, and robustness [1], [2]. To achieve such equivariance, one of the most fundamental ingredients is the *equivariant representation* of the environment.

In this work, we focus on the task of generating 6-DoF grasp poses for 3D objects. The primary objective is, given a dataset comprised of objects and their corresponding grasp poses, to train a model that can generate grasp poses for unseen, novel objects. The principle of $\mathrm{SO}(3)$-equivariance plays a pivotal role in the grasp pose generation task. As depicted in Figure 1, this principle dictates that, for any given rotated object, the generated grasp poses should be consistently rotated. This equivariance is vital for accurately training the generator, ensuring that it remains relevant and applicable irrespective of the object's orientation.

Recent studies have investigated the grasp pose generation task incorporating the equivariance. Zhu et al. [3] examined $\mathrm{SE}(2)$-equivariance within the context of planar (3-DoF) grasping, while Huang et al. [4] explored $\mathrm{SE}(3)$-equivariance in 6-DoF grasping scenarios. Despite these advancements, both approaches are constrained by their generation of a limited number of grasp poses. This limitation inherently increases the risk of grasping failures, which may arise from environmental obstacles or the robot's configuration constraints.

*Equal contribution. [1]Byeongdo Lim, Jongmin Kim, Jihwan Kim, and Frank C. Park are with the Department of Mechanical Engineering, Seoul National University. [2]Yonghyeon Lee is with the Center of AI and Natural Sciences (CAINS), Korea Institute for Advanced Study (KIAS). [3]Frank C. Park is with the SAIGE.

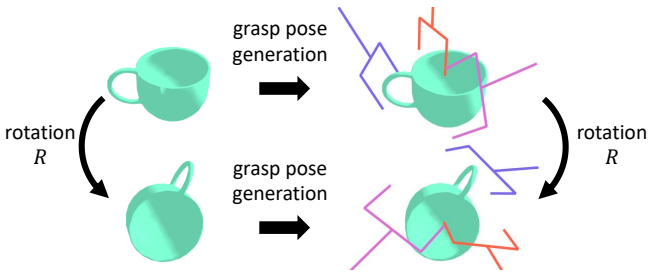

Fig. 1. An example of $\mathrm{SO}(3)$-equivariant grasp pose generation: the generated grasp poses consistently rotate with the object.

Learning the distribution of grasp poses can mitigate this issue by enabling the generation of multiple grasp poses from the learned distribution, rather than being confined to limited set. Furthermore, this approach facilitates the adoption of various grasping strategies for diverse downstream tasks. 6-DOF GraspNet [5] and $\mathrm{SE}(3)$-DiffusionFields [6] are proposed in this regard. However, these models do not fulfill the $\mathrm{SO}(3)$-equivariance requirement, resulting in less-than-desirable performance. The fundamental reason lies in their lack or only partial inclusion of a component required for learning equivariant representations of objects.

We propose a novel approach, termed **Equivariant Conditional Manifold Flows for grasping (GraspECMF)**, for learning the distribution of grasp poses that complies with the $\mathrm{SO}(3)$-equivariance. Leveraging $\mathrm{SO}(3)$-equivariant representations of objects extracted through vector neurons [7], we extend the equivariant manifold flow framework [8] – originally designed to learn invariant unconditional distributions on manifolds – to learn invariant *conditional* distributions specifically tailored for equivariant grasp pose generation. Our method builds upon three core ideas: (i) employing the objects' $\mathrm{SO}(3)$-equivariant representations as the conditional variables, (ii) extending the equivariant manifold flow framework to learn invariant conditional distributions, and (iii) designing a novel equivariant layer for our method.

Through experiments conducted using the ACRONYM dataset [9] and the Nvidia Isaac Gym simulator [10], we validate that our model surpasses existing state-of-the-art methods, 6-DOF GraspNet and SE(3)-DiffusionFields, in accurately learning the grasp pose distribution and generating grasp poses with a higher success rate.

II. METHOD

A. Learning Grasp Pose Distribution

The goal of the grasp pose generation is to find the distribution of 6-DoF grasp poses $x \in \mathrm{SE}(3)$ given the 3D point cloud $\mathcal{P}$ representing a geometry of an object. We train

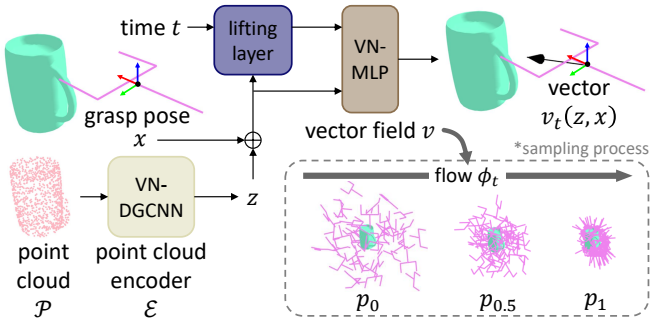

Fig. 2. The architecture of the equivariant conditional manifold flows utilized for equivariant grasp pose generation and the grasp pose sampling process. $\oplus$ denotes the concatenation of lists of 3D vectors.

a conditional generative model $q(x|\mathcal{P})$, approximated with a deep neural network, by a dataset $\mathcal{D} = \{\mathcal{P}_i, \{x_i^j\}_j | x_i^j \in \mathrm{SE}(3)\}_i$: each i-th object's point cloud $\mathcal{P}_i$ is labelled by multiple grasp poses $\{x_i^j\}_j$. We adopt encoder-based framework, which is standard to extract representations of $\mathcal{P}$ via an encoder $\mathcal{E}$, denoted by $z = \mathcal{E}(\mathcal{P}) \in \mathcal{Z}$. Then we model $q(x|\mathcal{P})$ as the distribution conditioned on z, such that $p(x|z) = p(x|\mathcal{E}(\mathcal{P})) = q(x|\mathcal{P})$ holds.

We employ the Continuous Normalizing Flow [11] as the generative model, as depicted in Figure 2. This model generates grasp poses by initially sampling from a simple prior distribution p_0 on the $\mathrm{SE}(3)$ manifold and pushing forward these samples via the flow ϕ_t, which is constructed from a time-dependent vector field v_t. Distinct from traditional flow models, our model utilizes the representation z as the conditional variable. Thus, the time-variant conditional vector field $v_t(z, x)$, is optimized to construct a conditional flows $\phi_t(z, x)$, such that the pushed-forward samples follow the target conditional distribution $p(x|z)$.

B. Equivariant Grasp Pose Generation with Equivariant Representations

In this section, we propose an $\mathrm{SO}(3)$-equivariant grasp pose generation method, utilizing $\mathrm{SO}(3)$-equivariant representations of objects. Consider the rotation group $\mathrm{SO}(3)$ and denote the group action of $g \in \mathrm{SO}(3)$ on $\mathcal{P}$ by $g \cdot \mathcal{P}$ and that on x by $g \cdot x$ (3D rotation of the points in $\mathcal{P}$ and the grasp pose x with respect to the reference frame). To ensure that the generated grasp poses are equivariant, conditional distribution $q(x|\mathcal{P})$ should be invariant: $q(g \cdot x | g \cdot \mathcal{P}) = q(x|\mathcal{P})$.

In particular, for the model involving the point cloud encoder $\mathcal{E}$ and its representation $z \in \mathcal{Z}$, ensuring the invariance of the conditional distribution $q(x|\mathcal{P})$ necessitates satisfying these conditions: (i) a group action of $\mathrm{SO}(3)$ is defined in $\mathcal{Z}$ with $g \cdot z$ denoting the action applied to z, (ii) an encoder should be $\mathrm{SO}(3)$-equivariant, i.e., $g \cdot z = \mathcal{E}(g \cdot \mathcal{P})$, and (iii) $p(x|z)$ should be $\mathrm{SO}(3)$-invariant, i.e., $p(x|z) = p(g \cdot x | g \cdot z)$. Then it achieves the invariance condition as follows:

$$\begin{aligned} q(g \cdot x | g \cdot \mathcal{P}) &= p(g \cdot x | \mathcal{E}(g \cdot \mathcal{P})) \\ &= p(g \cdot x | g \cdot z) = p(x|z) = q(x|\mathcal{P}). \end{aligned} \tag{1}$$

To construct the invariant distribution $p(x|z)$ conditioned on the representation, we extend the equivariant manifold flow framework proposed by Katsman et al. [8]. This framework is designed to learn an invariant unconditional distribution by utilizing an invariant prior distribution $p_0(x)$ and an equivariant vector field $v_t(x)$ which uniquely induces an equivariant flow $\phi_t(x)$. We extend this framework to learn an invariant *conditional* distribution, particularly by leveraging equivariant representations as the conditional variables. Therefore, we model the invariant conditional distribution $p(x|z)$ by using an equivariant conditional vector field $v_t(z, x)$ and an invariant prior distribution $p_0(x|z)$. Throughout, we use an unconditional prior distribution $p_0(x|z) = p_0(x)$ that is invariant, i.e., $p_0(g \cdot x) = p_0(x)$, for simplicity. We refer to our framework as Equivariant Conditional Manifold Flows for grasping (GraspECMF).

We train two neural networks. One is an equivariant neural vector field $v_t(z, x)$ to push forward the given invariant prior distribution $p_0(x)$ to model the target invariant distribution $p(x|z)$. The other is an equivariant encoder network $\mathcal{E}(\mathcal{P})$ to extract equivariant representations. Figure 2 illustrates the entire model and its grasp pose sampling process. The network details are explained in the subsequent section.

C. Network Implementation

We adopt vector neurons (VN) [7] to model both $v_t(z, x)$ and $\mathcal{E}(\mathcal{P})$. We use the standard VN-DGCNN to model $\mathcal{E}(\mathcal{P})$. On the other hand, it is not straightforward to directly use the architectures introduced in [7] – designed for 3D point cloud inputs – for our equivariant vector field $v_t(z, x)$. To use the VN architectures, the input should only consist of a list or set of 3D vectors. Consider the input variables of the vector field. The representation z obtained by the VN-DGCNN is a list of 3D vectors and $x \in \mathrm{SE}(3)$ can be considered as a list of 3D vectors. However, the time variable t is a scalar value, making it challenging to use the VN architectures.

In this work, we propose an *equivariant lifting layer* designed to elevate the scalar variable t to a list of vectors, enabling its integration with the VN architecture. Consider a list of scalar values represented as a column vector $s \in \mathbb{R}^{C_1 \times 1}$ – with the time variable corresponds to the case when $C_1 = 1$. And let $V \in \mathbb{R}^{C_2 \times 3}$ be a list of 3D vectors, e.g., in our case, it corresponds to (z, x) or its sub-list.

A lifting layer is a mapping

$$f_{\mathrm{lift}} : \mathbb{R}^{C_1 \times 1} \times \mathbb{R}^{C_2 \times 3} \to \mathbb{R}^{C_1 \times 3}, \tag{2}$$

and it is an equivariant layer if it satisfies $g \cdot f_{\mathrm{lift}}(s, V) = f_{\mathrm{lift}}(s, g \cdot V)$ for any $g \in \mathrm{SO}(3) \subset \mathbb{R}^{3 \times 3}$, where $g \cdot V$ for $V \in \mathbb{R}^{C \times 3}$ is defined by the matrix multiplication between V and g^T, i.e., $g \cdot V := V g^T$.

We construct the equivariant lifting layer as follows:

$$f_{\mathrm{lift}}(s, V) = s f_{\mathrm{equi}}(V), \tag{3}$$

where $f_{\mathrm{equi}} : \mathbb{R}^{C_2 \times 3} \to \mathbb{R}^{1 \times 3}$ is any equivariant mapping, i.e., $f_{\mathrm{equi}}(g \cdot V) = g \cdot f_{\mathrm{equi}}(V)$. It is trivial to show that this construction leads to the equivariance of f_{lift}. For f_{equi}, we use the VN architecture [7]. Finally, with the proposed equivariant lifting layer, we construct an equivariant neural networks to model the time-dependent vector field $v_t(z, x)$.

TABLE I

THE EVALUATION RESULTS OF EMD AND GRASP SUCCESS RATE

	EMD			**Grasp success rate (%)**		
	None	z	SO(3)	None	z	SO(3)
6-DOF GraspNet (VAE)	0.8788	0.8916	0.6040	19.27	17.48	51.68
6-DOF GraspNet (GAN)	0.8653	0.8707	0.7179	11.72	13.22	18.63
SE(3)-DiffusionFields	0.9458	0.8483	0.6269	13.02	22.50	**83.80**
GraspECMF (Ours)	**0.3490**	**0.3457**	**0.3634**	**81.45**	**84.03**	83.75

III. EXPERIMENTS

A. Experiment Settings

1) Dataset and Training: We utilize a dataset obtained from the Mug category of the ACRONYM dataset [9], which comprises 101 distinct mugs along with the poses of the Franka Panda gripper configured to grasp them. For the data augmentation, we constructed three strategies: *None* denotes no augmentation, *z-aug* denotes augmenting data by rotating through z-axis, *SO(3)-aug* denotes augmenting data by random arbitrary rotation in SO(3). For the test dataset, we utilize the dataset with the SO(3)-aug strategy. Riemannian Flow Matching [12] is employed to train our model.

2) Baselines: We compare our model with 6-DOF GraspNet [5] and SE(3)-DiffusionFields [6]. 6-DOF GraspNet comprises two versions utilizing Variational Autoencoder (VAE) and Generative Adversarial Network (GAN) architectures, respectively. SE(3)-DiffusionFields employs a diffusion model for generating the grasp poses.

3) Metrics: The evaluation metrics we utilize are Earth Mover's Distance (EMD) [13] and the grasp success rate. EMD measures the discrepancy between generated and ground-truth grasp poses, calculating the minimum distance (geodesic distance on the SE(3) manifold) necessary to align them. The grasp success rate is evaluated using the Nvidia Isaac Gym simulator [10], where the simulation involves grasping and shaking the object, determining success if the gripper holds the object afterward. We test 100 generated grasp poses per object, and both EMD and grasp success rate are averaged across three random rotations for each object.

B. Results

The evaluation results are presented in Table I and Figure 3. The column headers in the second row indicate the augmentation strategy of the training dataset used in each experiment. Figure 3 shows the grasp pose generation results on test data for the models trained with SO(3) augmentation.

The existing methods do not account for the SO(3)-equivariance in grasp pose generation, leading to insufficient performance with high EMD in both None and z-aug settings, as shown in Table I. Conversely, our GraspECMF incorporates SO(3)-equivariance, which results in better performance as indicated by low EMD in these settings. Moreover, even when arbitrary rotations augment the training dataset (SO(3)-aug), there exists a significant discrepancy in the EMD values between our GraspECMF and the baselines, underscoring our method's enhanced data efficiency and generalizability.

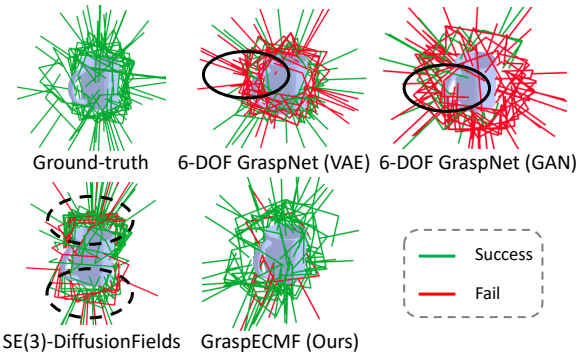

Fig. 3. Grasp pose generation results on test data for the models trained with SO(3) augmentation. 6-DOF GraspNet models produce ungraspable poses (black circles), while SE(3)-DiffusionFields focuses on specific parts (black dashed circles), which lack diversity in generating grasp poses.

Our method's superior performance is also evident in grasp success rates presented in Table I. 6-DOF GraspNet (VAE) and 6-DOF GraspNet (GAN) exhibit low success rates due to generating unfeasible grasp poses that collide with the cup, as indicated by black circles in Figure 3. SE(3)-DiffusionFields similarly shows low success rates in both None and z-aug settings. In contrast, our GraspECMF consistently achieves high success rates across all settings, benefiting from its incorporation of SO(3)-equivariance.

In the majority of scenarios, GraspECMF exhibits superior performance; however, it is marginally surpassed by SE(3)-DiffusionFields in the SO(3)-aug setting. The reason behind this phenomenon is shown in Figure 3. The grasp poses generated by GraspECMF are distributed across a diverse range of graspable areas on the cup, closely mirroring the distribution of the ground-truth grasp poses. In contrast, the grasp poses generated by SE(3)-DiffusionFields tend to be concentrated on specific regions of the mug, which is indicated by black dashed circles. Therefore, the comparable grasp success rate of the SE(3)-DiffusionFields to our GraspECMF can be said to be due to a lack of diversity in grasp poses. In contrast, our GraspECMF achieves both the diversity of the generated grasp poses and a high grasp success rate.

IV. CONCLUSIONS

In this paper, we introduce GraspECMF, an SO(3)-equivariant grasp pose generation model. Our approach revolves around three main ideas: (i) the utilization of SO(3)-equivariant representations of objects, (ii) the extension of the equivariant manifold flow framework for learning invariant conditional distributions, termed Equivariant Conditional Manifold Flows (ECMF), and (iii) the design of a novel equivariant layer for our method. Unlike existing grasp pose generation methods, our model ensures SO(3)-equivariance in generating grasp poses, resulting in enhanced data efficiency and generalizability. We verify the equivariance of the generated grasp poses and conduct quantitative evaluations against baselines, demonstrating the superior grasp pose generation performance of our model.

ACKNOWLEDGMENT

This work was supported in part by IITP-MSIT grant 2021-0-02068 (SNU AI Innovation Hub), IITP-MSIT grant 2022-0-00480 (Training and Inference Methods for Goal-Oriented AI Agents), KIAT grant P0020536 (HRD Program for Industrial Innovation), ATC+ MOTIE Technology Innovation Program grant 20008547, SRRC NRF grant RS-2023-00208052, SNU-AIIS, SNU-IAMD, SNU BK21+ Program in Mechanical Engineering, and SNU Institute for Engineering Research.

Yonghyeon Lee was the beneficiary of an individual grant from CAINS supported by a KIAS Individual Grant (AP092701) via the Center for AI and Natural Sciences at Korea Institute for Advanced Study.

REFERENCES

[1] B. Lee, Y. Lee, S. Kim, M. Son, and F. C. Park, "Equivariant motion manifold primitives," in *7th Annual Conference on Robot Learning*, 2023.

[2] S. Kim, B. Lim, Y. Lee, and F. C. Park, "Se (2)-equivariant pushing dynamics models for tabletop object manipulations," in *Conference on Robot Learning*. PMLR, 2023, pp. 427–436.

[3] X. Zhu, D. Wang, O. Biza, G. Su, R. Walters, and R. Platt, "Sample efficient grasp learning using equivariant models," *arXiv preprint arXiv:2202.09468*, 2022.

[4] H. Huang, D. Wang, X. Zhu, R. Walters, and R. Platt, "Edge grasp network: A graph-based se (3)-invariant approach to grasp detection," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 3882–3888.

[5] A. Mousavian, C. Eppner, and D. Fox, "6-dof graspnet: Variational grasp generation for object manipulation," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 2901–2910.

[6] J. Urain, N. Funk, J. Peters, and G. Chalvatzaki, "Se (3)-diffusionfields: Learning smooth cost functions for joint grasp and motion optimization through diffusion," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 5923–5930.

[7] C. Deng, O. Litany, Y. Duan, A. Poulenard, A. Tagliasacchi, and L. J. Guibas, "Vector neurons: A general framework for so (3)-equivariant networks," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 12 200–12 209.

[8] I. Katsman, A. Lou, D. Lim, Q. Jiang, S. N. Lim, and C. M. De Sa, "Equivariant manifold flows," *Advances in Neural Information Processing Systems*, vol. 34, pp. 10 600–10 612, 2021.

[9] C. Eppner, A. Mousavian, and D. Fox, "Acronym: A large-scale grasp dataset based on simulation," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 6222–6227.

[10] V. Makoviychuk, L. Wawrzyniak, Y. Guo, M. Lu, K. Storey, M. Macklin, D. Hoeller, N. Rudin, A. Allshire, A. Handa, *et al.*, "Isaac gym: High performance gpu-based physics simulation for robot learning," *arXiv preprint arXiv:2108.10470*, 2021.

[11] R. T. Chen, Y. Rubanova, J. Bettencourt, and D. K. Duvenaud, "Neural ordinary differential equations," *Advances in neural information processing systems*, vol. 31, 2018.

[12] R. T. Chen and Y. Lipman, "Riemannian flow matching on general geometries," *arXiv preprint arXiv:2302.03660*, 2023.

[13] A. Tanaka, "Discriminator optimal transport," *Advances in Neural Information Processing Systems*, vol. 32, 2019.