Robotic Manipulation Learning with Equivariant Descriptor Fields: Generative Modeling, Bi-equivariance, Steerability, and Locality

Jiwoo Kim*[1] Hyunwoo Ryu*[1] Jongeun Choi[1,2]
Joohwan Seo[2] Nikhil Potu Surya Prakash[2] Ruolin Li[2] Roberto Horowitz[2]

[1]Machine Learning and Control Systems, Yonsei University, [2]University of California, Berkeley, (*Equal Contribution)

Abstract

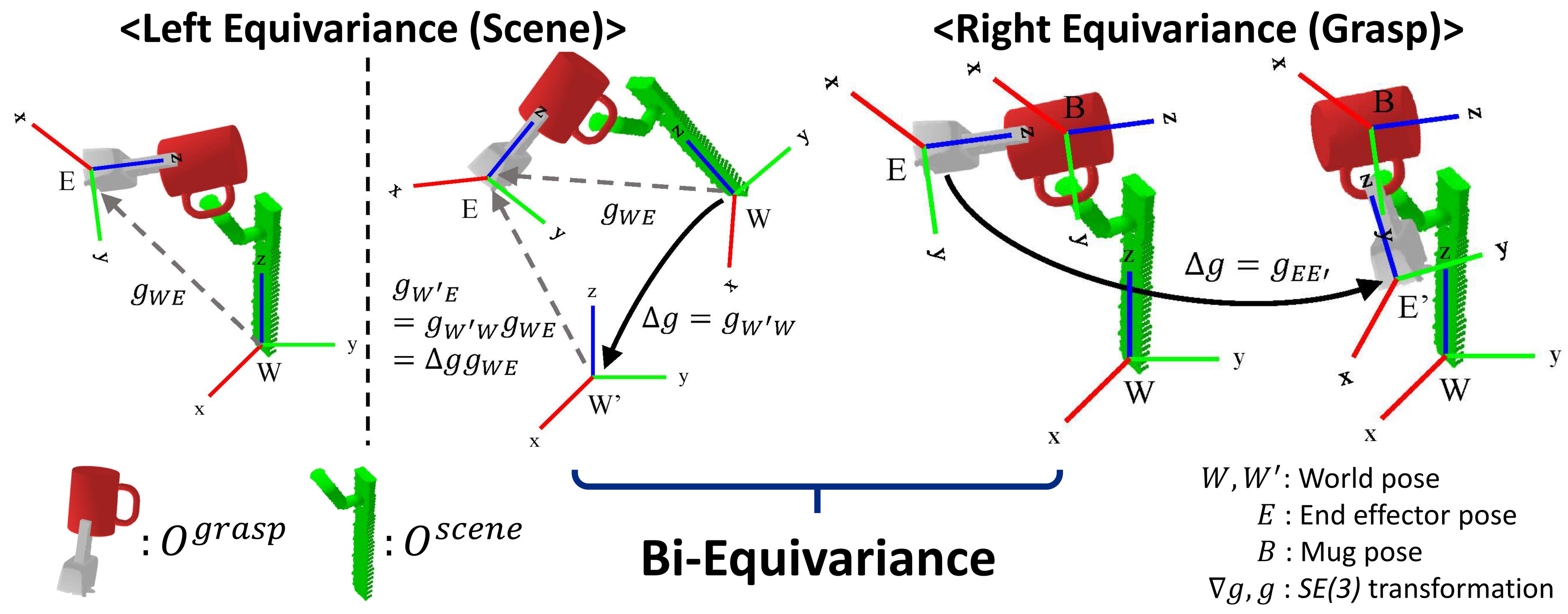
- We examine the design principles of recently proposed *Equivariant Descriptor Fields* (EDFs), highlighting the importance of four key concepts: **Generative Modeling**, **Bi-equivariance**, **Steerability**, and **Locality.**
- *Equivariant Descriptor Fields* (EDFs) are fully $SE(3)$-equivariant visual robotic manipulation models that can be **end-to-end trained from scratch** with **only 5~10 demos.**

Generative Modeling

- Expert demonstrations for manipulation are **mostly multimodal.** e.g. humans can pick the mug by **the rim** or by **the handle**.
- Generative models are successful in learning the **proper multimodalities.**
- EDFs utilize energy-based model (EBM) approach to model the policy distribution, enabling both **end-to-end training** and **sampling**.

$$P(g|O^{scene}, O^{grasp}) = \frac{\exp[-\mathrm{E}(g|O^{scene}, O^{grasp})]}{\int_{g\in SE(3)} dg \exp[-\mathrm{E}(g|O^{scene}, O^{grasp})]}$$

Bi-equivariance

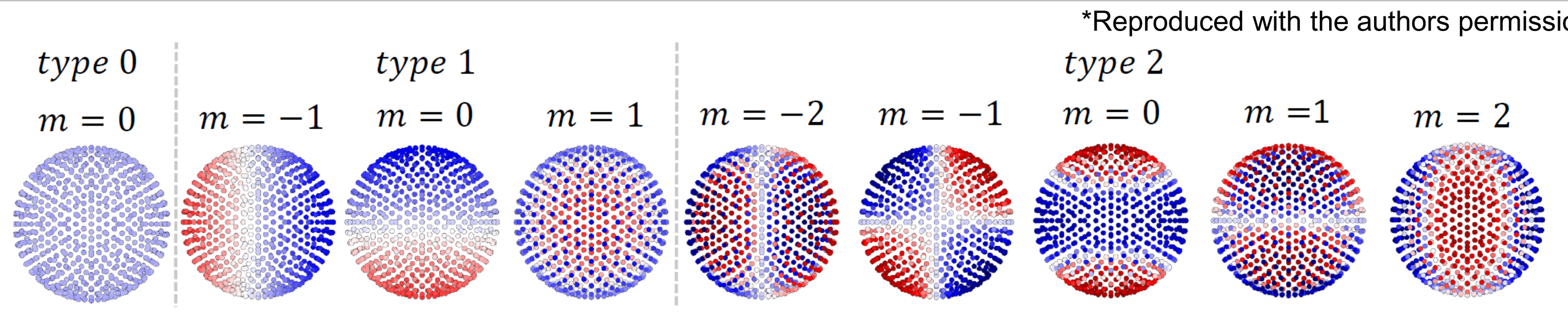


$$P(\Delta g g|\Delta g O^{scene}, O^{grasp}) = P(g|O^{scene}, O^{grasp}) = P(g\Delta g^{-1}|O^{scene}, O^{grasp}\Delta g)$$

- For successful pick-and-place manipulation, the model needs to generalize to the unseen pose of the object within the scene and the grasp, which may **significantly deviate from the trained demonstrations**.
- The model should be able to utilize the "**scene equivariance**" or "**left equivariance**" to adapt to unseen configurations of the target object in the scene.
- To be able to generalize to out-of-distribution grasps, it is necessary for the model to compensate for changes of the grasped object's pose through "**grasp equivariance**" or "**right equivariance**".
- "**Bi-equivariance**" combines the principles of left and right equivariance, enhancing generalizability and robustness under diverse configuration.

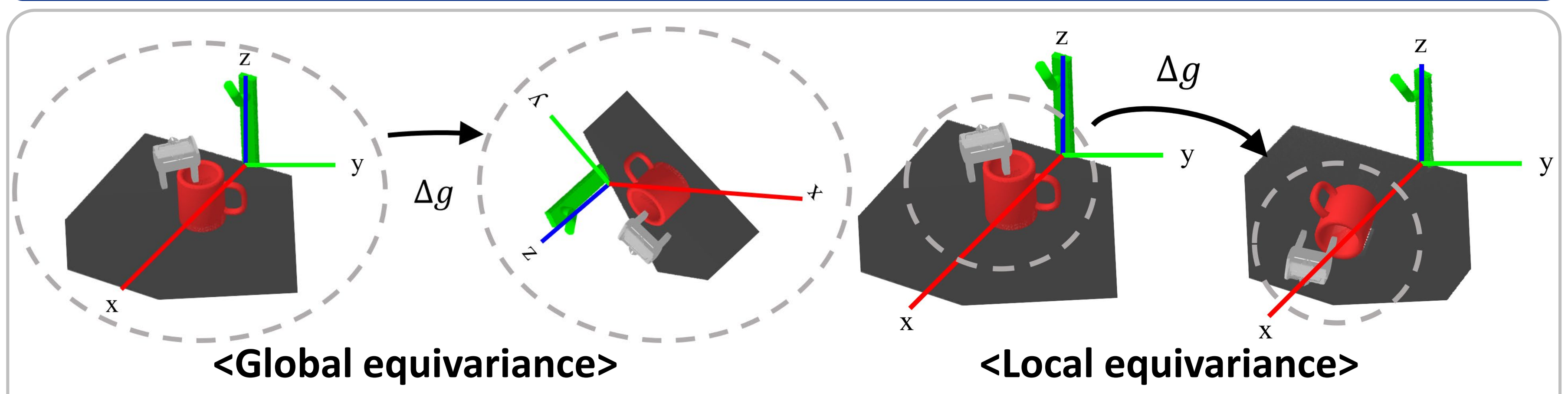
Steerable Representation

*Reproduced with the authors permission [3]



- According to the representation theory of the *SO(3)*-group, every equivariant vectors can be decomposed and categorized into *type-l* ($l = 0, 1, 2, \ldots$) *vectors.*
- **SE(3)-equivariant vector field of type-0** is rotationally invariant such that $f(gx) = f(x)$. **Type-1 or higher SE(3)-equivariant vector fields** are rotationally equivariant such that $f(gx) = D_l(R)f(x)$ where $D_l(R)$ is the Wigner D-matrix of degree-l that *steers* type-l feature vectors.
- Steerable representations are highly effective at capturing the orientations of the **local geometries,** due to their **orientational sensitivity.**

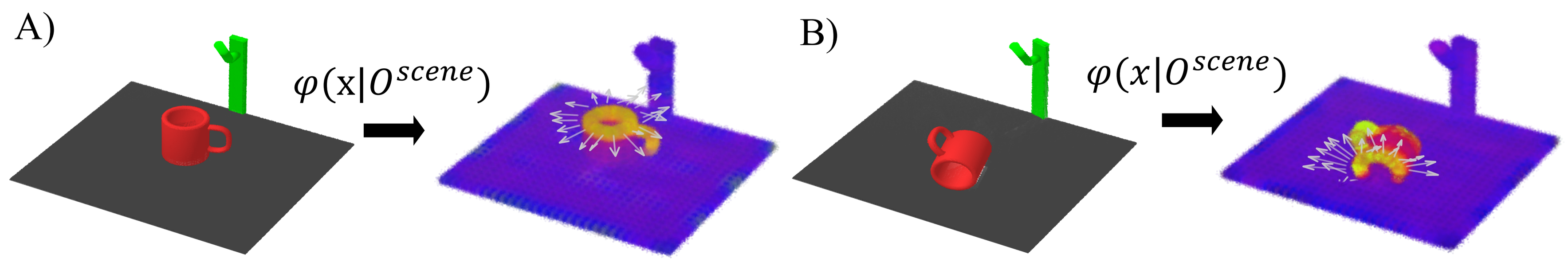
Locality



- Locality enhances **generalizability** by learning the shared local **geometric structure** of the target object.
- Locality removes the need for **object segmentation pipeline** for the input.

Equivariant Descriptor Fields

Colors: Type-0 (Rotation invariant) **Arrows**: Type-1 (Rotation equivariant) *Type-2 and higher are not visualized



- An equivariant descriptor field $\varphi(\cdot\,|O)$ generated by an input point cloud O is an $SE(3)$-equivariant vector field on $\mathbb{R}^3$ such that

$$\varphi(\Delta g x|\Delta g O) = D(R)\varphi(x|O)$$

x: position, O : point cloud, $\forall \Delta g = (p, R) \in SE(3)$

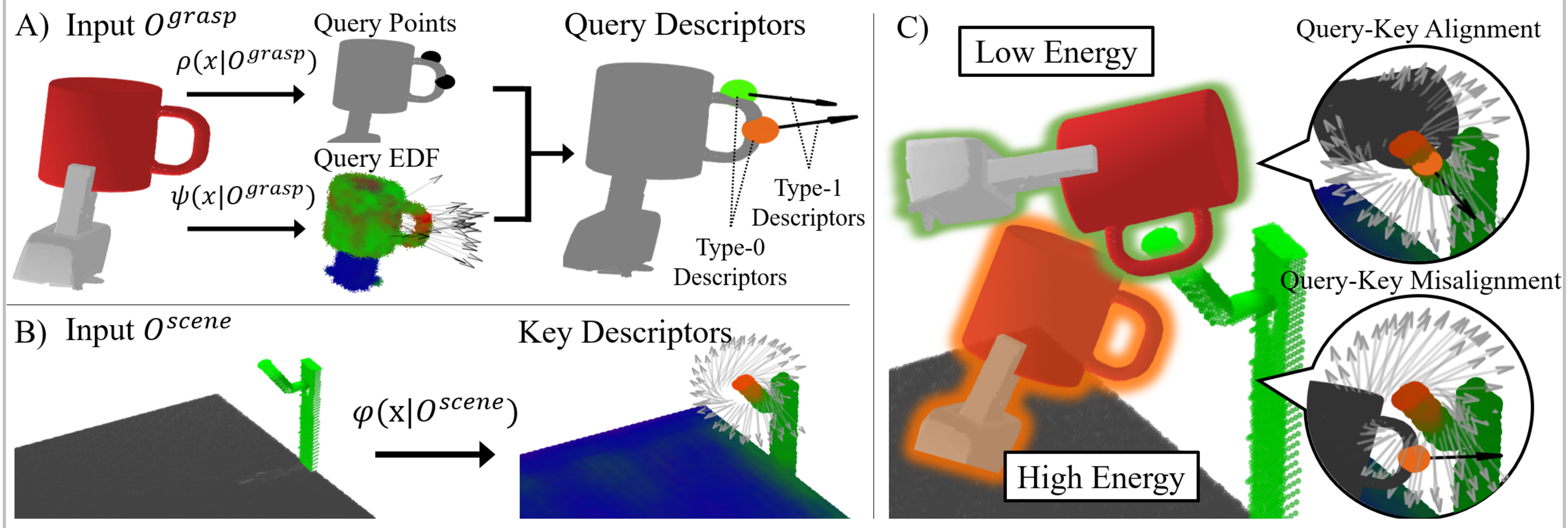
- By exploiting the steerability of the EDFs, the **bi-equivariant energy function** can be constructed as follows:

$$\mathrm{E}(g|O^{scene}, O^{grasp}) = \int_{\mathbb{R}^3} d^3x \rho(x|O^{grasp}) ||\varphi(gx|O^{scene}) - D(R)\psi(x|O^{grasp})||^2$$

$\psi_\theta(x|O^{grasp})$: Query−EDF, $\varphi_\theta(x|O^{scene})$: Key−EDF, $\rho_\theta(x|O^{grasp})$: Equivariant Query Density

- For the energy function to be **tractable**, the query density is modeled as weighted query points composed of weighted sum of 3D Dirac delta function:

$$\rho_\theta(x|O^{grasp}) = \sum_{i=1}^{N_q} w_\theta(q_{i;\theta}(O^{grasp})|O^{grasp})\delta^{(3)}(x - q_{i;\theta}(O^{grasp}))$$

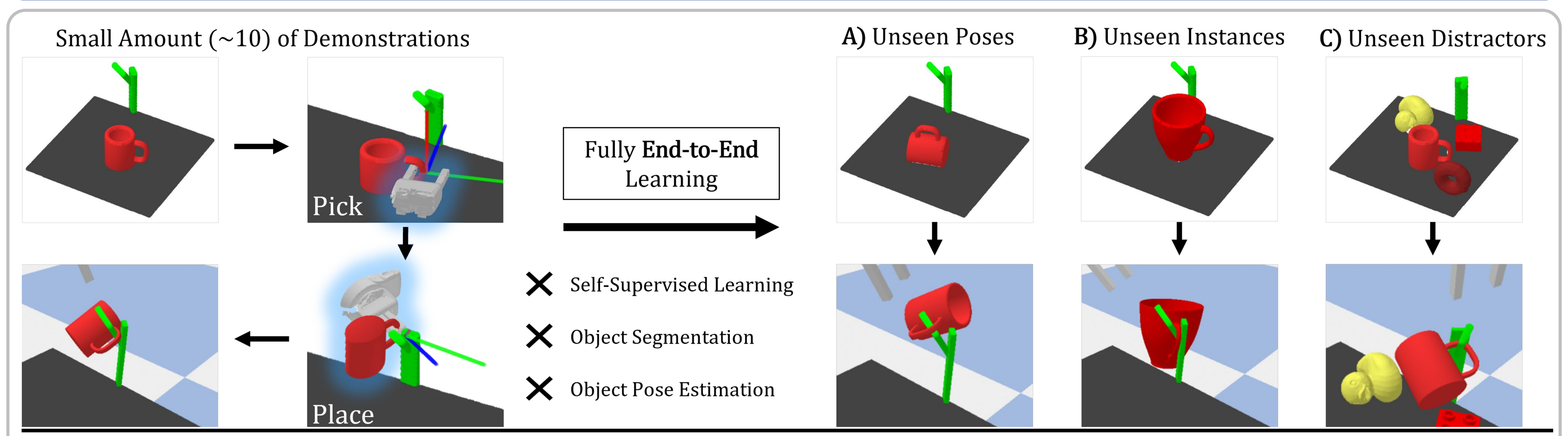


A) The **query EDF** generated from grasp point cloud O^{grasp}, assigns the **query descriptors** to the **query points**.

B) Similarly, the **key EDF** is generated from O^{scene}.

C) The energy values are computed by matching the transformed query descriptors to the key descriptors. The **lower energy** case has a better **alignment of the query and the key descriptors**, meaning it has **higher probability**. MCMC methods are used to sample end-effector poses according to their energy.

Experiment Results



	Mug			Bowl			Bottle		
	Pick	Place	**Total**	Pick	Place	**Total**	Pick	Place	**Total**
Unseen Instances									
$SE(3)$-TNs	1.00	0.36	0.36	0.76	1.00	0.76	0.20	1.00	0.20
EDFs (Ours)	1.00	**0.97**	**0.97**	**0.98**	1.00	**0.98**	**1.00**	1.00	**1.00**
Unseen Poses									
$SE(3)$-TNs	0.00	N/A	0.00	0.00	N/A	0.00	0.00	N/A	0.00
EDFs (Ours)	**1.00**	1.00	**1.00**	**1.00**	1.00	**1.00**	**0.95**	1.00	**0.95**
Unseen Distracting Objects									
$SE(3)$-TNs	1.00	0.63	0.63	1.00	1.00	1.00	0.96	0.92	0.88
EDFs (Ours)	1.00	**0.98**	**0.98**	1.00	1.00	1.00	**0.99**	**1.00**	**0.99**
Unseen Instances, Arbitrary Poses & Distracting Objects									
$SE(3)$-TNs	0.25	0.04	0.01	0.09	1.00	0.09	0.26	0.88	0.23
EDFs (Ours)	**1.00**	**0.95**	**0.95**	**0.95**	1.00	**0.95**	**0.95**	**1.00**	**0.95**

	Mug			Bowl			Bottle		
Descriptor Type	Pick	Place	**Total**	Pick	Place	**Total**	Pick	Place	**Total**
NDF-like (Type-0 Only)									
Inference Time	5.7s	8.6s	14.3s	6.1s	9.9s	16.0s	5.8s	17.3s	23.0s
Success Rate	0.84	0.77	0.65	0.60	0.95	0.57	0.66	0.95	0.63
EDFs (Type-0~3)									
Inference Time	5.1s	8.3s	13.4s	5.2s	10.4s	15.6s	5.2s	11.5s	16.7s
Success Rate	**1.00**	**0.95**	**0.95**	**0.95**	**1.00**	**0.95**	**0.95**	**1.00**	**0.95**

- EDFs evaluate the pick-and-place success rate for three different scenarios (mug hanging, bowl/bottle placing).
- EDFs are trained **from scratch with only ten demonstrations** for each scenario, using **no pre-training** or **object segmentation** pipelines.
- EDFs achieves >95% success rate even if previously **unseen target object instance** is provided in **unseen pose** with **unseen distracting objects**.
- EDFs outperform baselines ($SE(2)$-equivariant baseline [1], and type-0 only baseline [2]) by a significant margin in success rate.

[1] Andy Zeng et al., Transporter networks: Rearranging the visual world for robotic manipulation, CoRL 2020.
[2] Anthony Simeonov, Yilun Du et al., Neural descriptor fields: SE(3)-equivariant object representations for manipulation, ICRA 2022.
[3] Evangelos Chatzipantazis, Stefanos Pertigkiozoglou et al., SE(3)-Equivariant Attention Networks for Shape Reconstruction in Function Space, ICLR 2023

Original EDF Paper (ICLR 2023)
Workshop Paper (RSS 2023)