A Review of Diffusion Fields

Jaime Canizales

City University of New York

jaime.canizales@hunter.cuny.edu

April 26, 2024

Overview

Introduction

Problem Statement

How to solve the multi-objective optimization problem of pick and place, where a robot arm can pick up an object and place it somewhere else.

Why is this hard?

- The robot must find a grasp pose for the object it must pick up.
- The robot must generate a path in order to pick up the object, by taking into consideration its joint and velocity limits and the robot kinematics to accomplish a smooth trajectory(motion optimization).
- The robot must consider collision avoidance during trajectory generation(often approximated by learning based approaches).

Why is this hard?(cont.)

- Most solutions decouple the trajectory generation from the grasp pose estimation(For example what we did with swindepose).
- This can be problematic because of the disconnected systems. You may infer a grasp pose(there are many possible grasp poses) that has no possible trajectory generation solution in your second system.

Introduction(cont.)

Solution proposed in paper

Learning a SE(3) diffusion model for 6DOF grasping, giving rise to a novel framework for joint grasp and motion optimization without needing to decouple grasp selection from trajectory generation.

How is this done?

- Modify the trajectory generation function(which is a sum of cost functions), by adding another term for the grasp pose selection in cost function form.
- Use a diffusion model to model the cost function for grasp pose selection.
- solve for the gradient in the Lie algebra space because very hard otherwise.

Technical Information

- LIE Algebra
- Diffusion
- Motion Optimization

Inventor of LIE Algebra

Lie in 1896

- Marius Sophus Lie (17 December 1842 – 18 February 1899) was a Norwegian mathematician. He largely created the theory of continuous symmetry and applied it to the study of geometry and differential equations. He also made substantial contributions to the development of algebra.

LIE Algebra Introduction

- A lie algebra is a lie group at the identity element.
- A lie group is a group thats also a smooth(differentiable) manifold(locally similar to a euclidean space).
- A group is a set with an operation that satisfies the following constraints: the operation is associative and has an identity element, and every element of the set has an inverse element.
- A lie algebra has an operation called the lie bracket that satisfies the jacobi identity.

LIE Algebra Introduction(cont.)

- $SE(3)$ stands for the special euclidean group in 3 dimension is defined as the set of points that satisfy a point $\mathbf{H}=\begin{bmatrix} R & t \\ 0 & 1 \end{bmatrix} \in SE(3)$ where $R \in SO(3)$ and $t \in \mathbb{R}^3$.
- $SO(3)$ Stands for the special orthogonal group, and represents the set of orthogonal matrices that have determinant 1(these are all rotation matrices).

LIE Algebra

- The Lie Algebra is the tangent manifold space to the identity element $I = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$ and is isomorphic to $\mathbb{R}^6$ in which we can apply linear algebra(equations are applied in $\mathbb{R}^6$).

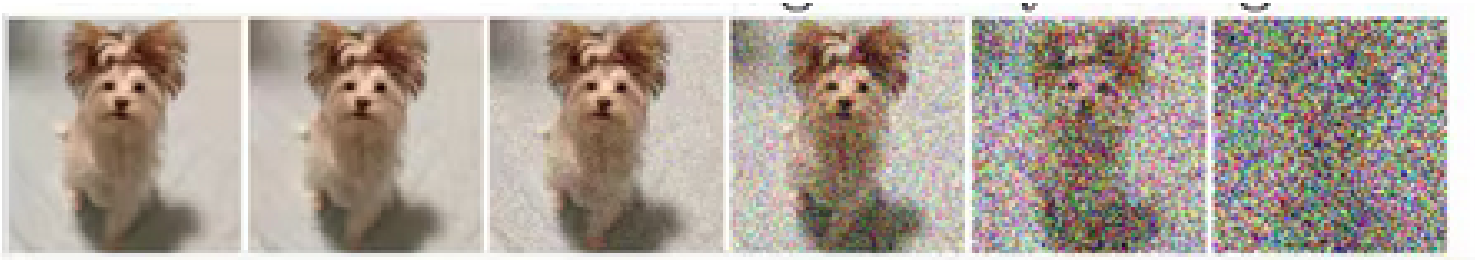
LIE Algebra (cont.)

A Gaussian distribution on LIE groups

$q(H|H_{\mu,}\Sigma) \propto exp(-0.5||Logmap(H_{\mu}^{-1}H)||^2_{\Sigma^{-1}})$

- Logmap : $SE(3) \rightarrow \mathfrak{se}(3)$(lie algebra) $\rightarrow \mathbb{R}^6$
- $H_{\mu} \in SE(3)$ is the mean.
- $\Sigma \in \mathbb{R}^{6\times6}$ is the covariance matrix.

Diffusion



- In physics diffusion is the movement of particles move from an area of high concentration to an area of low concentration until equilibrium is reached.
- Perturb the data distribution with $\rho_D(x)$ with Gaussian noise on L noise scales $\mathcal{N}(0, \sigma_k^2 I)$ with $\sigma_1 < \sigma_1 < ... < \sigma_L$.
- To obtain the noise perturbed distribution:
  $q_{\sigma_k}(\hat{x}) = \int_x \mathcal{N}(\hat{x}|x, \sigma_k I)\rho_D(x)$ where $\hat{x} = x + \epsilon$, $\epsilon \sim \mathcal{N}(0, \sigma_k I)$

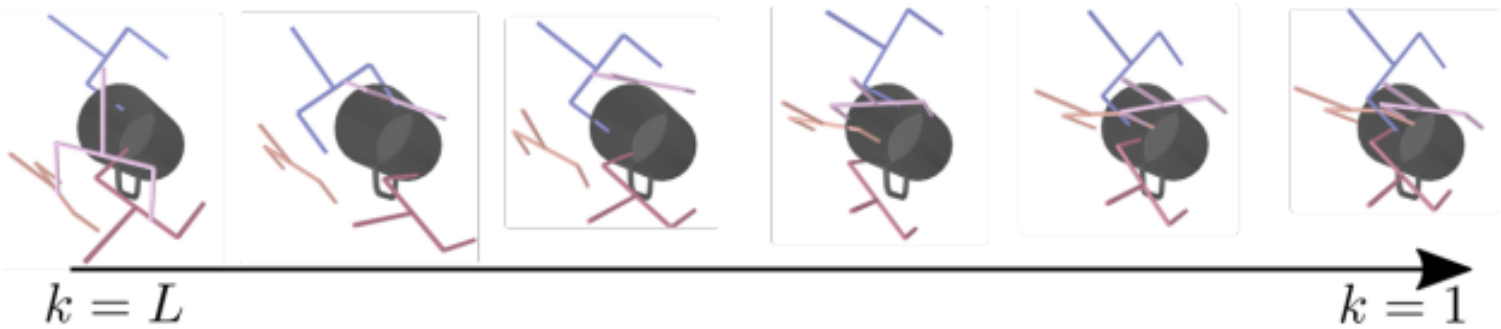
Denoising Score Matching

- estimate the score function(gradient of the log-likelihood) by training the vector field $s_\theta(x, k)$: $s_\theta(x, k) \approx \nabla_x log q_{\sigma_k}(x)$ for all k=1,...,L
- Thus, the training objective of DSM is:

$$\mathcal{L}_{dsm} = \frac{1}{L} \sum_{k=0}^{L} \mathbb{E}_{x,\hat{x}}[||s_\theta(x, k) - \nabla_{\hat{x}} log \mathcal{N}(\hat{x}|x, \sigma_k^2 I)||]$$

- $x \sim \rho_D(x)$ and $\hat{x} \sim \mathcal{N}(x, \sigma_k I)$

Annealed Langevin Markov Chain Monte Carlo



- Draw sample from the distribution $x_L \sim p_L(x)$ and then simulate the inverse langevian diffusion process for L steps, from $k = L$ to $k = 1$
- $x_{k-1} = x_k + \frac{\alpha_k^2}{2} s_\theta(x_k, k) + \alpha\epsilon, \epsilon \sim \mathcal{N}(0, I)$

From Euclidean Diffusion to SE(3) Diffusion

- A diffusion model in SE(3) is a vector field that outputs a vector $v \in \mathbb{R}^6$, for an arbitrary query point $H \in SE(3)$.
- $v = s_\theta(H, k)$ with scalar conditioning variable k determining the current noise scale.

DSM in SE(3) Step 1

- Generate a perturbed data point in SE(3) from the distribution $\hat{H} = q(\hat{H}|H, \sigma_k I)$
- Where $\hat{H} = H * Expmap(\epsilon), \epsilon \sim \mathcal{N}(0, \sigma_k^2 I)$
- Expmap: $\mathbb{R}^6 \rightarrow SE(3)$

DSM in SE(3) Step 2

- Train:

$$\mathcal{L}_{dsm} = \frac{1}{L}\sum_{k=0}^{L}[||s_\theta(\hat{H}, k) - \frac{Dlogq(\hat{H}|H, \sigma_k I)}{DH}||]$$

- Sampling:

$$H_{k-1} = Expmap(\frac{\alpha_k^2}{2}s_\theta(H, k) + \alpha_k \epsilon)H_k$$

Architecture

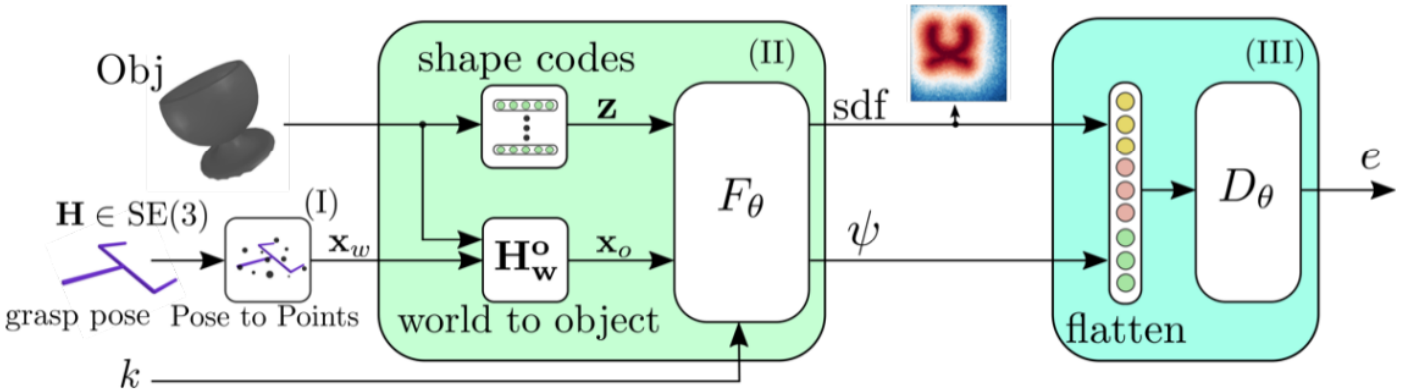


Fig. 3: SE(3)-DiF's architecture for learning 6D grasp pose distributions. We train the model to jointly learn the objects' sdf and to minimize the denoising loss. Given grasp pose $\boldsymbol{H} \in \mathrm{SE}(3)$ we transform it to a set of 3D points $\boldsymbol{x}_w \in \mathbb{R}^{N\times 3}$ (**I**). Next, we transform the points into the object's local frame, using the object's pose $\boldsymbol{H}^o_w$. Given the resulting points $\boldsymbol{x}_o$ and the object's shape code $\boldsymbol{z}$ we apply the feature encoder F_θ (**II**) to obtain a object and grasp-related features (sdf, ψ)$\in \mathbb{R}^{N\times(\psi+1)}$. Finally, (**III**) we flatten the features and compute the energy e through the decoder D_θ. We provide a point-cloud-based

- shape codes are the known shapes of the object in training data
- sign distance field(sdf) - A function that takes a position as input and outputs the distance from from that position to the nearest part of the shape(supervised learning pipeline).

Grasp and motion optimization with diffusion models

- Given a trajectory $\{q_t\}_t^T = 1$ consisting of T way points
- Using Energy based models (EBM), we model our score function $s_\theta(H, k) = \frac{-DE_\theta(H,k)}{DH} = c_n$ to add it to the motion optimization function.
- Aim to minimize trajectory:

$$\tau^* = argmin_\tau \sum_j w_j c_j(\tau)$$

- Finally, we can sample trajectories:

$$\tau_{k-1} = \tau_k + \frac{\alpha_k^2}{2} \nabla_{\tau_k} logq(\tau|k) + \alpha\epsilon, \epsilon \sim \mathcal{N}(0, I)$$

Training

- Training for SE(3)-DiF as a 6DoF grasp pose generative model using the Acronym dataset. This simulation-based dataset contains successful 6DoF grasp poses for a variety of objects from ShapeNet. It focuses on the collection of successful grasp poses for 90 different mugs (approximately 90K 6DoF grasp poses).

Tests

1. Generate a set of grasp poses from the learned models and evaluate on successful grasping and diversity.
2. Evaluate the quality of the trained model when used as an additional cost term for grasp and motion optimization. Compare the performance of solving a grasp and motion optimization problem jointly , w.r.t. the state-of-the-art approaches that decouple the grasp selection and motion planning, or heuristically combine them.
3. Validate the performance of the method in a set of real robot experiments.

Test 1 (Evaluation of 6DoF grasp pose generation)

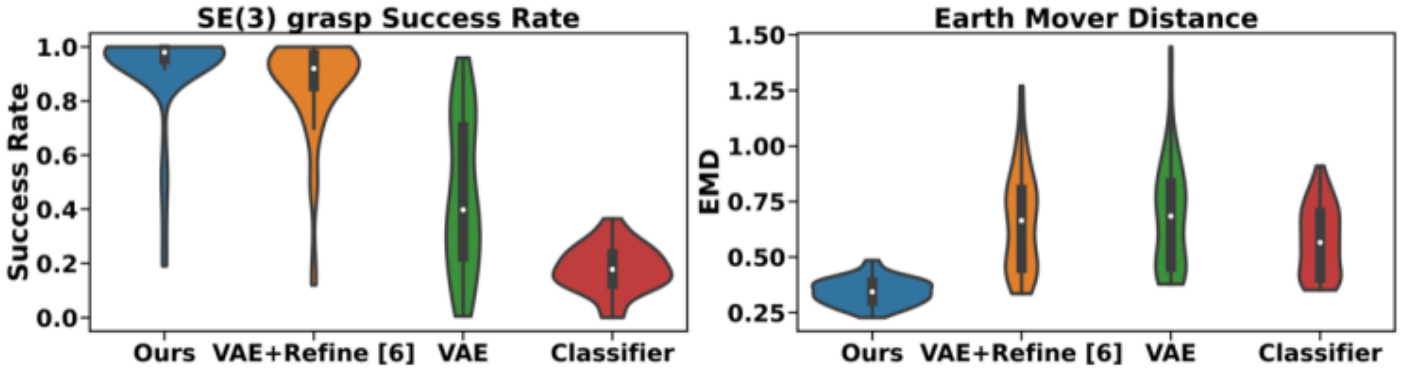


Fig. 4: 6D grasp pose generation experiment. Left: Success rate evaluation. Right: Earth Mover Distance (EMD) evaluation metrics (lower is better).

- Consider 90 different mugs and evaluate 200 generated grasps per mug.
- The Earth Mover Distance(EMD) measures the divergence between two empirical probability distributions.

Test 2 (Performance on grasp and motion optimization)

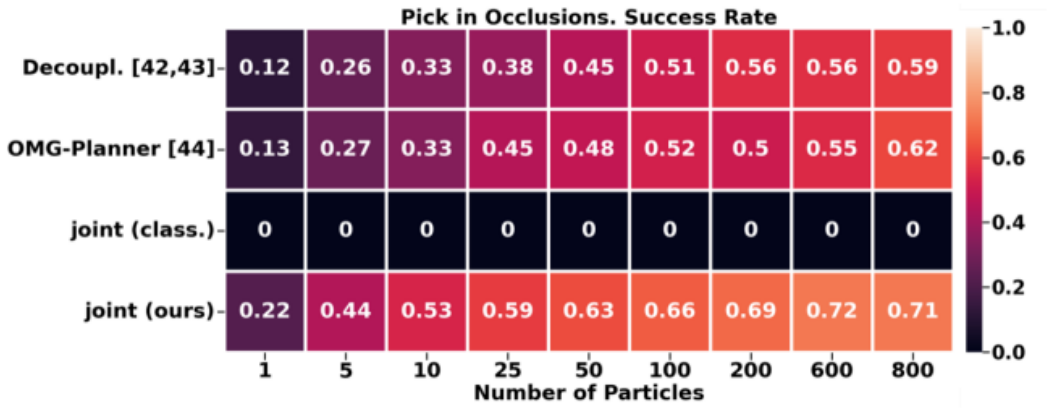


Fig. 5: Evaluation Pick in occlusion. We measure the success rate of 4 different methods based on different number of initializations.

- The success is measured based on the robot being able to grasp the object at the end of the execution
- Generate the trajectories by integrating our learned grasp SE(3)-DiF as an additional cost function to the motion optimization objective function

Test 2(cont.)

- Then, given a set of initial trajectory samples, obtained from a Gaussian distribution with a block diagonal matrix, we apply gradient descent methods, to iteratively improve the trajectories on the objective function.
- Requires 25 particles(trajectories) to match the success rate of the de- coupled approach with 800 particles

Test 3 (Grasp and motion optimization on real robots)

- the robot has to pick up a mug from various poses in a scene without any clutter, achieved 100% (20 successes / 20 trials) pickup-success.
- upside down mugs with 90% (18/20) success
- picking in occludes scenes with 95% (19/20) success
- having to pick and place the mug in a desired pose inside the shelf of Fig. 1 with 100% (20/20) success.
- did not suffer from sim2real discrepancies.

Conclusion

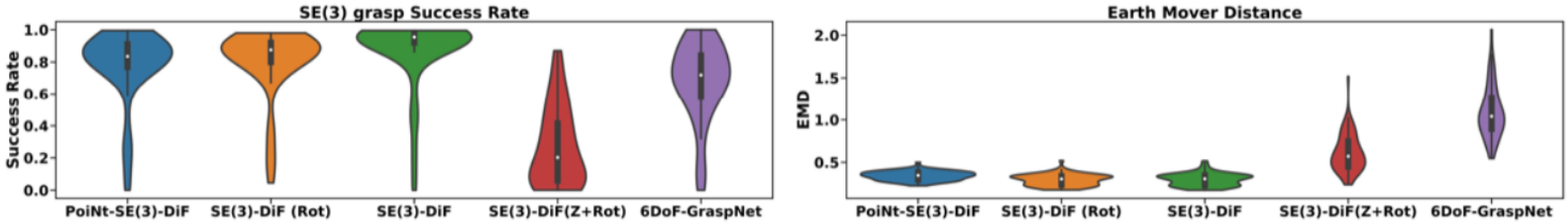


Fig. 12: Evaluation of the Success for picking with occlusions. PoiNt-SE(3)-DiF refers to the model with a pointcloud encoder, SE(3)-DiF (Rot) to the model in which the pose is infer from the pointcloud, SE(3)-DiF the model in which both the pose and shape are known, SE(3)-DiF (Z+Rot) the model in which both the object pose and shape codes are inferred from the pointcloud, and 6DoF-Graspnet [6].

- Comparison to state of the art at the time.

References

Julen Urain, Niklas Funk, Jan Peters, Georgia Chalvatzaki (2023)

SE(3)-DiffusionFields: Learning smooth cost functions for joint grasp and motion optimization through diffusion