Confronting Ambiguity in 6D Object Pose Estimation via Score-Based Diffusion on SE(3)

Tsu-Ching Hsiao, Hao-Wei Chen, Hsuan-Kung Yang, and Chun-Yi Lee

Elsa Lab, Department of Computer Science, National Tsing Hua University, Hsinchu City, Taiwan

CVPR SEATTLE, WA JUNE 17-21, 2024

Abstract

Addressing pose ambiguity in 6D object pose estimation from a single RGB images presents significant challenge, particularly due to object symmetries or occlusions, as illustrated in Fig. 1. Although state-of-the-art regression methods use a symmetry-aware loss, this approach requires symmetry annotations obtained through extensive manual labor and time. In contrast, non-parametric methods model the pose uncertainty as a distribution across $SO(3)$, eliminating the need for symmetry annotations. However, exhaustive grid search is required for training and sampling. In response, we propose a novel score-based diffusion models operating on $SE(3)$, which overcomes aforementioned problem. As depicted in Fig. 3, the pose is gradually refined from an initial guess throughout the denoising process.

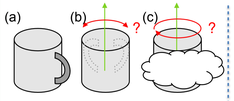

Figure 1. An illustration of pose ambiguity. (a) Unambiguous, (b) self-occlusion, and (c) occlusion.

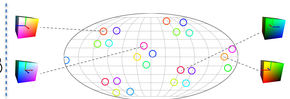

Figure 2. Modeling pose ambiguity as a distribution on $SO(3)$.

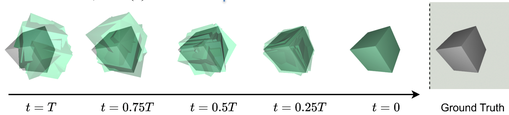

Figure 3. Visualization of the denoising process of our score-based diffusion model on $SE(3)$ for 6D object pose estimation.

Background

Lie groups & Parametrization of SE(3)

A Lie group $\mathcal{G}$ and its associate Lie algebra $\mathfrak{g}$ are related through mappings: $\mathrm{Exp} : \mathfrak{g} \to \mathcal{G}, \mathrm{Log} : \mathcal{G} \to \mathfrak{g}$. This work considers two groups $R^3SO(3)$ and $SE(3)$, each with a different parametrization and composition rule:

- $R^3SO(3)$:
  Parametrization: $\langle \mathbb{R}^3, SO(3) \rangle$
  Composition rule: $(R_2, T_2)(R_1, T_1) = (R_2R_1, T_2 + T_1)$
- $SE(3)$:
  Parametrization: $(R, T) = (\mathrm{Exp}(\phi), \mathbf{J}_l(\phi)\rho)$
  Composition rule: $(R_2, T_2)(R_1, T_1) = (R_2R_1, T_2 + R_2T_1)$

Methodology

Given an RGB image I that displays the object of interest, our goal is to estimate the 6D object poses $X = (R, T) \in SE(3)$. This can be interpreted as sampling poses from a pose distribution $X \sim p(X \mid I)$, which captures the inherent pose uncertainty. We model the distribution using score-based pose diffusion model.

Score-Based Diffusion Model on Lie Group

Gaussian perturbation kernel in Lie group defined as:

$$p_\Sigma(Y \mid X) := \mathcal{N}_{\mathcal{G}}(Y; X, \Sigma) \triangleq \frac{1}{\zeta(\Sigma)} \exp\left(-\frac{1}{2}\mathrm{Log}(X^{-1}Y)^\top \Sigma^{-1} \mathrm{Log}(X^{-1}Y)\right)$$

Sampling $\tilde{X}$ from $\mathcal{N}_{\mathcal{G}}(\tilde{X}; X, \sigma)$ is achieved by:

$$\tilde{X} = X\mathrm{Exp}(z),\ z \sim \mathcal{N}(0, \sigma^2 I), \text{ where } z \in \mathfrak{se}(3)$$

The score of the perturbation kernel:

$$\nabla_{\tilde{X}} \log p_\sigma(\tilde{X} \mid X) = -\frac{1}{\sigma^2}\mathbf{J}_r^{-\top}(z)z$$

Denoising Score Matching objective:

$$\mathscr{L}(\theta; \sigma) \triangleq \frac{1}{2}\mathbb{E}_{p_{\text{data}}(X)}\mathbb{E}_{\tilde{X} \sim \mathcal{N}_{\mathcal{G}}(X, \Sigma)}\left[\left\| s_\theta(\tilde{X}, \sigma) - \nabla_{\tilde{X}} \log p_\sigma(\tilde{X} \mid X) \right\|_2^2\right]$$

Denoising process (geodesic random walk):

$$\tilde{X}_{i+1} = \tilde{X}\mathrm{Exp}(\epsilon_i s_\theta(\tilde{X}_i, \sigma_i) + \sqrt{2\epsilon_i} z_i),\ z_i \sim \mathcal{N}(0, I) \quad (1)$$

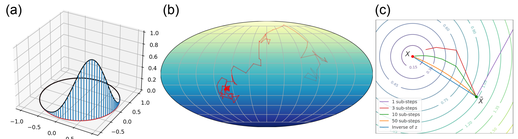

Figure 4. Visualization of (a) Gaussian kernel on $SO(2)$, (b) geodesic random walk on $SO(3)$ (Mellow projection), and (c) a denoising step from $\tilde{X}$ to X.

Efficient Computation of the Stein Score

The score can be simplified as:

$$\nabla_{\tilde{X}} \log p_\sigma(\tilde{X} \mid X) = -\frac{1}{\sigma^2}z$$

if $\mathcal{G}$ satisfies the following condition:

$$\mathbf{J}_l(z) = \mathbf{J}_r^\top(z),\ \mathbf{J}_l^{-1}(z) = \mathbf{J}_r^{-\top}(z),\ \text{and}\ \mathbf{J}_l(z)z = z$$

The Stein score on $SO(3)$ and $R^3SO(3)$ can be simplified as they satisfy the condition. However, $SE(3)$ does not possess this property as we can prove:

$$\mathbf{J}_r^{-\top}(z) = \left(\mathbf{J}_l^{-1}(-z)\right)^\top = \begin{bmatrix} \mathbf{J}_l^{-1}(\phi) & 0 \\ \mathbf{Z}(\rho, \phi) & \mathbf{J}_l^{-1}(\phi) \end{bmatrix} \neq \mathbf{J}_l^{-1}(z)$$

This inequality indicates the discrepancy between the score vector and the denoising direction, which impede the convergence of the reverse process.

To address this problem, we observe that the denoising direction is the continuous integral of score vectors along the path as presented in Fig. 4. Thus, we propose a surrogate Stein score for training on $SE(3)$, defined as :

$$\tilde{s}_{\tilde{X}}(\tilde{X}, \sigma) \triangleq -\frac{1}{\sigma^2}z$$

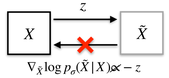

Figure 5. Discrepancy between the score vector and the denoising direction.

The Proposed Framework

We present our framework in Fig. 5. The conditioning module generates condition variable c from images and diffusion time steps, which is used to guide the denoising process. The denoising module estimate scores $s_\theta(\tilde{x}_i, \sigma_i)$ of a noisy pose $\tilde{x}_i$ and iteratively refine it with Eq. (1).

For the condition operation, we propose Fourier-based conditioning mechanism to capture the inherent periodic features of $SO(3)$ space, defined as:

$$f_i(x, c) = \sum_{j=0}^{d-1} \mathbf{W}_{ij}\left(\mathbf{A}_j(c)\cos(\pi x_j) + \mathbf{B}_j(c)\sin(\pi x_j)\right)$$

Scale-based conditioning: $f(x, c) = \mathbf{A}(c)x + B(c)$

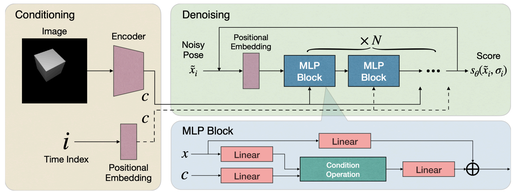

Figure 5. Framework overview.

References

[1] Kieran et. al. Implicit-PDF: Nonparametric Representation of Probability Distributions on the rotation manifold. ICML, 2021.
[2] Tomáš et. al. T-LESS: An RGB-D Dataset for 6D Pose Estimation of Texture-less Objects. WACV, 2017.
[3] Tomáš et. al. BOP Challenge 2020 on 6D Object Localization. ECCV Workshops, 2020.

Experimental Results

Quantitative Results on SYMSOL [1]

Table 1. The baselines utilize ResNet50 as the backbone. We report the average angular distances in degrees.

Methods	SYMSOL (Spread in degrees ↓)					
	Avg.	tet.	cube	icosa.	cone	cyl.
DBN	22.44	16.70	40.70	29.50	10.10	15.20
Implicit-PDF	3.96	4.60	4.00	8.40	1.40	1.40
HyperPosePDF	1.94	3.27	2.18	3.24	0.55	0.48
Normalizing Flows	0.70	0.60	0.60	1.10	0.50	0.50
Ours (ResNet34)	0.42	0.43	0.44	0.52	**0.35**	0.35
Ours (ResNet50)	**0.37**	**0.28**	**0.32**	**0.40**	0.53	**0.31**

Quantitative Results on SYMSOL-T

Table 2. We developed SYMSOL-T based on SYMSOL by adding random translations. We use ResNet34 as the backbone, and report the average angular distances in degrees for rotation R and the average distances for translation t.

Methods	SYMSOL-T (Spread in degrees ↓)									
	tet.		cube		icosa.		cone		cyl.	
	R	t	R	t	R	t	R	t	R	t
Regression	2.92	0.064	2.86	0.05	2.46	0.037	1.84	0.058	2.24	0.049
Iterative regression	4.25	0.048	4.2	0.037	29.33	0.026	1.63	0.037	2.34	0.032
Ours ($R^3SO(3)$)	1.38	0.017	1.93	**0.010**	29.35	**0.009**	1.33	**0.016**	0.86	**0.010**
Ours ($SE(3)$)	**0.59**	**0.016**	**0.58**	0.011	**0.64**	0.012	**0.54**	0.016	**0.41**	0.011

Quantitative Results on T-LESS [2]

Table 3. We use ResNet34 as the backbone. We report the three metrics from the BOP challenge [3], rotation errors within 2, 5, 10 degrees, and translation errors within 0.02, 0.05, 0.1 unit.

Methods	T-LESS (Accuracy % ↑)								
	MSPD	MSSD	VSD	R@2	R@5	R@10	T@2	T@5	T@10
GDRNPP	90.17	75.06	67.60	21.60	71.18	90.56	90.31	96.09	98.10
Ours ($R^3SO(3)$)	85.73	52.03	48.41	27.98	72.42	89.26	60.37	79.75	89.62
Ours ($SE(3)$)	**93.16**	60.17	56.88	**47.21**	**86.94**	**94.78**	71.72	92.03	97.15

Inference Time Analysis on T-LESS

Table 4. We assess inference time across different denoising steps (step skipping) on the T-LESS dataset.

Methods	Steps	Inference time	FPS	MSPD	MSSD	VSD
Ours ($R^3SO(3)$)	100	0.041	24	85.73	52.03	48.41
	50	0.021	47	85.46	52.18	48.41
	10	0.005	188	85.57	52.25	48.77
	5	0.003	307	85.67	53.11	49.59
Ours ($SE(3)$)	100	0.050	20	93.16	60.17	56.88
	50	0.026	38	93.00	59.96	56.64
	10	0.006	161	92.79	60.35	57.08
	5	0.004	250	92.40	59.30	56.15

Visualization

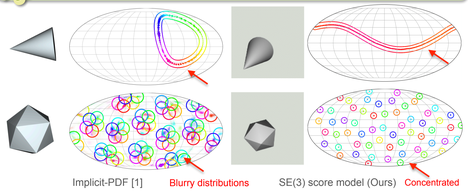

Figure 6. Visualization of Implicit-PDF [1] and our $SE(3)$ models.

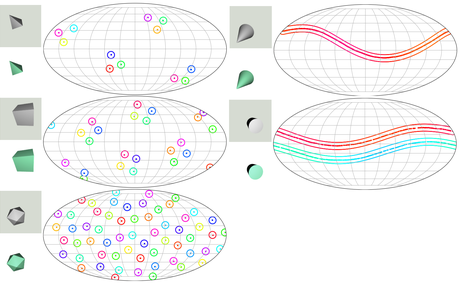

Figure 7. Visualization of our $SE(3)$ score model on SYMSOL-T.

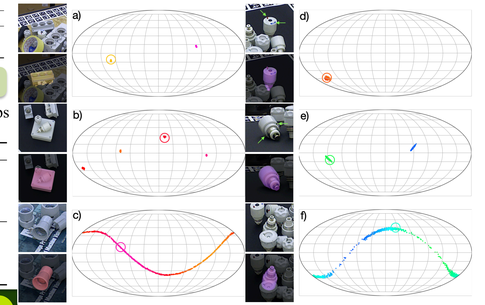

Figure 8. Visualization of our $SE(3)$ score model on T-LESS.

Conclusion

1. We presented a novel approach that applies diffusion models to the $SE(3)$ for addressing the pose ambiguity issue.
2. We developed the SYMSOL-T dataset, which enrich the SYMSOL dataset with randomly sampled translations.
3. Our experiments confirmed the applicability of our $SE(3)$ score model in solving pose ambiguity, and demonstrated its efficacy in real-world application.

Contact & Acknowledgement

Questions?
joehsiao@gapp.nthu.edu.tw
cylee@cs.nthu.edu.tw

NSTC TWCC TAIWAN COMPUTING CLOUD