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, National Tsing Hua University
{joehsiao, jaroslaw1007, hellochick}@gapp.nthu.edu.tw
cylee@cs.nthu.edu.tw

Abstract

Addressing pose ambiguity in 6D object pose estimation from single RGB images presents a significant challenge, particularly due to object symmetries or occlusions. In response, we introduce a novel score-based diffusion method applied to the $SE(3)$ group, marking the first application of diffusion models to $SE(3)$ within the image domain, specifically tailored for pose estimation tasks. Extensive evaluations demonstrate the method's efficacy in handling pose ambiguity, mitigating perspective-induced ambiguity, and showcasing the robustness of our surrogate Stein score formulation on $SE(3)$. This formulation not only improves the convergence of denoising process but also enhances computational efficiency. Thus, we pioneer a promising strategy for 6D object pose estimation.

1. Introduction

Estimating the six degrees of freedom (DoF) pose of objects from a single RGB image remains a formidable task, primarily due to the presence of ambiguity induced by symmetric objects and occlusions. Symmetric objects exhibit identical visual appearance from multiple viewpoints, whereas occlusions arise when key aspects of an object are concealed either by another object or its own structure. This can complicate the determination of its shape and orientation. Pose ambiguity presents a unique challenge as it transforms the direct one-to-one correspondence between an image and its associated object pose into a complex one-to-many scenario, which can potentially leads to significant performance degradation for methods reliant on one-to-one correspondence. Despite extensive exploration in the prior object pose estimation literature [10, 19, 21, 38, 39], pose ambiguity still remains a persisting and unresolved issue.

Recent advancements in pose regression have introduced the use of symmetry-aware annotations to improve pose estimation accuracy [38, 42, 58, 62]. These methods typically employ symmetry-aware losses that can tackle the pose ambiguity problem. The efficacy of these losses, nevertheless, depend on the provision of symmetry annotations, which can be particularly challenging to obtain for objects with intricate shapes or under occlusion. An example is a textureless cup, where the true orientation becomes ambiguous if the handle is not visible. The manual labor and time required to annotate the equivalent views of each object under such circumstances is impractical.

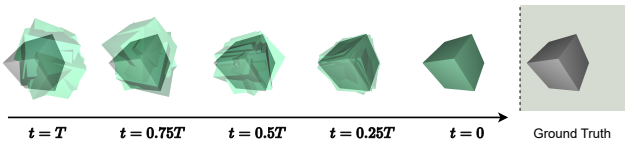

Figure 1. Visualization of the denoising process of our score-based diffusion method on $SE(3)$ for 6DoF pose estimation.

Several contemporary studies have sought to eliminate the reliance on symmetry annotations by treating 'equivalent poses' as a multi-modal distribution, reframing the original pose estimation problem as a density estimation problem. Methods such as Implicit-PDF [39] and HyperPose-PDF [23] leverage neural networks to implicitly characterize the non-parametric density on the rotation manifold $SO(3)$. While these advances are noteworthy, they also introduce new complexities. For instance, the computation during training requires exhaustive sampling across the whole $SO(3)$ space. Moreover, the accuracy of inference is dependent on the resolution of the grid search, which necessitates a significant amount of grid sampling. These computational limitations are magnified when extending to larger spaces such as $SE(3)$ due to the substantial memory requirements.

Recognizing these challenges, the research community is pivoting towards diffusion models (DMs) [16, 54–56], which are effective in handling multi-modal distributions. Their effectiveness lies in the iterative sampling process, which incorporates noises and enables a more focus exploration of the pose space while reducing computational demands. As diffusion models refrain from explicit den-

sity estimation, this property enables them to handle large spaces and high-dimensional distributions. In prior endeavors, the authors in [28, 32] applied the denoising diffusion probabilistic model (DDPM) [16] and score-based generative model (SGM) [56] to the $SO(3)$ rotation manifold, effectively recovering unknown densities on the $SO(3)$ space. On the other hand, other research efforts [59, 68] have extended the application of diffusion models to the more complex $SE(3)$ space, which enlightens the potential applicability of diffusion models in object pose estimation tasks.

In light of the above motivations, we introduce a novel approach that applies diffusion models to the $SE(3)$ group for object pose estimation tasks, specifically aimed at addressing the pose ambiguity problem. This method draws its inspiration from the correlation observed between rotation and translation distributions, a phenomenon often resultant from the perspective effect inherent in image projection. We propose that by jointly estimating the distribution of rotation and translation on $SE(3)$, we may secure more accurate and reliable results as shown in Fig. 1. To the best of our knowledge, this is the first work to apply diffusion models to $SE(3)$ within the context of image space. To substantiate our approach, we have developed a new synthetic dataset, called SYMSOL-T, based on the original SYMSOL dataset [39]. It enhances the original dataset with randomly sampled translations, offering a more rigorous testbed to evaluate our method's effectiveness in capturing the joint density of object rotations and translations.

Following the motivations discussed above, we have extensively evaluated our $SE(3)$ diffusion model using the synthetic SYMSOL-T dataset and a real-world T-LESS [20] dataset. The experimental results affirm the model's competence in handling $SE(3)$, which successfully addresses the pose ambiguity problem in 6D object pose estimation. Moreover, the $SE(3)$ diffusion model has proven effective in enhancing rotation estimation accuracy and robustness. Importantly, the surrogate Stein score formulation we propose on $SE(3)$ exhibits improved convergence in the denoising process compared to the score calculated via automatic differentiation. This not only highlights the robustness of our method, but also demonstrates its potential to handle complex dynamics in object pose estimation tasks.

2. Background

2.1. Lie Groups and Their Applications

A Lie group, denoted by $\mathcal{G}$, serves as a mathematical structure with broad applicability due to its dual nature as both a group and a smooth (or differentiable) manifold. The latter is a topological space that can be locally approximated as a linear space. In accordance with the axioms governing groups, a composition operation is formally defined as a mapping $\circ : \mathcal{G} \times \mathcal{G} \to \mathcal{G}$. The composition operation, along with the associated inversion map, exhibits smoothness properties consistent with the group structure. For notational convenience in subsequent analyses, the composition of two group elements $X, Y \in \mathcal{G}$ is succinctly denoted as $X \circ Y = XY$. Every Lie group $\mathcal{G}$ has an associated Lie algebra, denoted as $\mathfrak{g}$. A Lie group and its associated Lie algebra are related through the following mappings: $\text{Exp} : \mathfrak{g} \to \mathcal{G}$, $\text{Log} : \mathcal{G} \to \mathfrak{g}$. In the context of pose estimation, two Lie groups are commonly employed: $SO(3)$ and $SE(3)$. The Lie group $SO(3)$ and its associated Lie algebra $\mathfrak{so}(3)$ can represent rotations in three-dimensional Euclidean space. On the other hand, the Lie group $SE(3)$, along with its corresponding Lie algebra $\mathfrak{se}(3)$, can be employed to describe rigid-body transformations, which incorporate both rotational and translational elements in Euclidean space. Such group structures form the mathematical basis for analyzing and solving complex problems, especially for six Degrees of Freedom (6DoF) pose estimation.

2.2. Lie Group Representation of Transformations

A variety of parametrizations for these transformation groups are discussed in [53]. This work considers two types of transformation groups, each characterized by a distinct manifold structure and the accompanying parametrizations: $R^3SO(3)$ and $SE(3)$. The former parametrization, which segregates rotations $R \in SO(3)$ and translations $T \in \mathbb{R}^3$ into a composite manifold $\langle \mathbb{R}^3, SO(3) \rangle$, denotes its Lie algebra as $\langle \mathbb{R}^3, \mathfrak{so}(3) \rangle$. $R^3SO(3)$ employs a composition rule defined by $(R_2, T_2)(R_1, T_1) = (R_2R_1, T_2 + T_1)$. This parametrization, which is prevalent in several prior diffusion models on $R^3SO(3)$ due to its simplicity as discussed in [59, 68], induces a separate diffusion process for both R and T. Another parametrization, $SE(3)$, formulates elements within the Lie algebra as $\tau = (\rho, \phi) \in \mathfrak{se}(3)$, wherein ρ and ϕ correspond to infinitesimal translations and rotations at the identity element's tangent space, respectively. The corresponding group elements within $SE(3)$ are represented as $(R, T) = (\text{Exp}(\phi), \mathbf{J}_l(\phi)\rho)$, where $\mathbf{J}_l$ denotes the left-Jacobian of $SO(3)$. The composition rule for the $SE(3)$ parametrization is expressed as $(R_2, T_2)(R_1, T_1) = (R_2R_1, T_2 + R_2T_1)$. The integration of both rotations and translations within $SE(3)$ gives rise to a diffusion process that emulates the elaborate dynamics of rigid-body motion.

2.3. Score-Based Generative Modeling

Consider independent and identically distributed (i.i.d.) samples $\{\mathbf{x}_i \in \mathbb{R}^D\}_{i=1}^N$ drawn from a data distribution $p_{\text{data}}(\mathbf{x})$. The *(Stein) score* of a probability density $p(\mathbf{x})$ is the gradient of its logarithm, denoted as $\nabla_{\mathbf{x}} \log p(\mathbf{x})$ [27]. In the framework of score-based generative models (SGMs), an important formulation within the spectrum of diffusion models, data undergo a gradual transformation toward a known prior distribution. Such a

distribution is often selected for computational tractability [61], and this process is termed the *forward* process. The forward process is characterized by a series of increasing noise levels $\{\sigma_i\}_{i=1}^{L}$, which are ordered such that $\sigma_{\text{min}} = \sigma_1 < \sigma_2 < \ldots < \sigma_L = \sigma_{\text{max}}$. The selection of σ_{min} and σ_{max} as sufficiently small and large values respectively facilitates the approximation of $p_{\sigma_{\text{min}}}(\mathbf{x})$ to $p_{\text{data}}(\mathbf{x})$ and of $p_{\sigma_{\text{max}}}(\mathbf{x})$ to the Gaussian distribution $\mathcal{N}(\mathbf{x}; \mathbf{0}, \sigma_{\text{max}}^2\mathbf{I})$. This process utilizes a perturbation kernel $p_\sigma(\tilde{\mathbf{x}}|\mathbf{x}) = \mathcal{N}(\tilde{\mathbf{x}}; \mathbf{x}, \sigma^2\mathbf{I})$, and the perturbed distribution is given by $p_\sigma(\tilde{\mathbf{x}}) = \int p_{\text{data}}(\mathbf{x}) p_\sigma(\tilde{\mathbf{x}}|\mathbf{x}) d\mathbf{x}$. In the Noise Conditional Score Network (NCSN) [55], a network $s_{\boldsymbol{\theta}}(\mathbf{x}, \sigma)$ parameterized by θ is trained to estimate the score via a Denoising Score Matching (DSM) objective [61] as follows:

$$\begin{aligned} \boldsymbol{\theta}^* &= \arg\min_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}; \sigma) \\ &\triangleq \frac{1}{2}\mathbb{E}_{p_{\text{data}}(\mathbf{x})}\mathbb{E}_{\tilde{\mathbf{x}}\sim\mathcal{N}(\mathbf{x},\sigma^2 I)}\left[\|s_{\boldsymbol{\theta}}(\tilde{\mathbf{x}}, \sigma) - \nabla_{\tilde{\mathbf{x}}} \log p_\sigma(\tilde{\mathbf{x}}|\mathbf{x})\|_2^2\right]. \end{aligned} \tag{1}$$

The optimal score-based model $s_{\boldsymbol{\theta}^*}(\mathbf{x}, \sigma)$ aims to match $\nabla_{\mathbf{x}} \log p(\mathbf{x})$ as closely as possible across the entire range of σ values in the set $\{\sigma_i\}_{i=1}^{L}$. During the sample generation phase, score-based generative models employ an iterative *reverse* process. Specifically, in the context of the Noise Conditional Score Network (NCSN), the Langevin Markov Chain Monte Carlo (MCMC) method is utilized to execute M steps. This process is designed to produce samples in a sequential manner from each $p_{\sigma_i}(\mathbf{x})$, expressed as follows:

$$\tilde{\mathbf{x}}_i^m = \tilde{\mathbf{x}}_i^{m-1} + \epsilon_i s_{\boldsymbol{\theta}^*}(\tilde{\mathbf{x}}_i^{m-1}, \sigma_i) + \sqrt{2\epsilon_i}\mathbf{z}_i^m, \quad m = 1, 2, ..., M, \tag{2}$$

where $\epsilon_i > 0$ denotes the step size, and $\mathbf{z}_i^m$ represents a standard normal variable. Overall, diffusion based models, especially SGMs, provide a solid framework for handling complex data distributions. They serve as the foundation for the denoising procedure employed by our methodology.

3. Related Work

3.1. Methodologies for Dealing with Pose Ambiguity

Non-probabilistic modeling. In the realm of object pose estimation, pose ambiguity remains a significant challenge, often stemming from an object that exhibits identical visual appearances from different perspectives [38]. A variety of strategies have been explored in the literature to directly address this issue, including the application of symmetry supervisions and point matching algorithms [1, 64]. Regression-based approaches, such as those presented in [11, 31, 58, 62], aim to minimize pose discrepancy by selecting the closest candidate within a set of ambiguous poses. Some researchers [44, 46], on the other hand, introduce constraints to the regression targets (especially regarding rotation angles) to mitigate ambiguity. Moreover, certain approaches [25, 42, 63] suggest regressing to a predetermined set of geometric features derived from symmetry annotations. These prior arts often necessitate manual annotations of equivalent poses and are limited in dealing with other sources of pose ambiguities, such as those caused by occlusion and self-occlusion [38].

Probabilistic modeling. On the other hand, several studies have investigated methods to model the inherent uncertainty in pose ambiguity. This involves the quantification and representation of uncertainty associated with the estimated poses. Some works have employed parametric distributions such as Bingham distributions [10, 12, 41] and von-Mises distributions [45, 69] to model orientation uncertainty. Other approaches, such as in [37], utilize normalizing flows [48] to model distributions within rotational space. A number of studies [23, 30, 39] employ non-parametric distributions to implicitly represent rotation uncertainty on $SO(3)$. These methods primarily focus on modeling distributions on $SO(3)$, leaving the joint distribution modeling of rotation and translation unexplored.

3.2. Diffusion Probabilistic Models and Their Application Domains

Diffusion models on Euclidean space. Diffusion probabilistic models [16, 54–56, 66] represent a class of generative models designed to learn the underlying probability distribution of data. They have been applied to various generative tasks, and have shown impressive results in several application domains, including image [2, 3, 7, 47, 49–51], video [17, 18, 67], audio [26, 65], and natural language processing [13, 34]. In the realm of human pose estimation, diffusion models have also been found useful in addressing joint location ambiguity, which arises from the projection of 2D keypoints into 3D space [9, 24].

Diffusion models on non-Euclidean space. To accommodate data residing on a manifold, the authors in [5] extended diffusion models to Riemannian manifolds, and leveraged Geodesic Random Walk [29] for sampling. Other studies [28, 32] applied the Denoising Diffusion Probabilistic Models (DDPM) [16] and score-based generative models [55, 56] to the $SO(3)$ manifold to recover the density of data on $SO(3)$. Further extensions of diffusion models have been attempted for tasks such as unfolding protein structures [68] and arm manipulations [59]. These approaches typically used $R^3SO(3)$ parametrization, which treated rotation and translation as separate entities for diffusion.

3.3. Diffusion Models on Lie Groups

Diffusion models on Lie groups have been explored in a range of applications [28, 32, 59, 68]. Nevertheless, these implementations vary in their choices of distributions and computational methods, which lead to diverse outcomes and different levels of computational efficiency. Table 1 presents a comparison of several previous diffusion model approaches along with our own. It highlights the distinct

Table 1. Comparison of different methods. $\triangle$ means closed form but with approximation. $\mathcal{N}_{SE(3)}$ please refer to Eq. (3).

Baselines	Group	Distribution	Closed Form	Diffusion Method	Diffusion Space	App. Domain
Leach *et al.* [32]	$SO(3)$	$IG_{SO(3)}$	✗	DDPM	$SO(3)$	Vector
Jagvaral *et al.* [28]	$SO(3)$	$IG_{SO(3)}$	✗	Score / Autograd	$SO(3)$	Vector
Urain *et al.* [59]	$R^3SO(3)$	$\mathcal{N}_{\mathbb{R}^3} \times \mathcal{N}_{SO(3)}$	✓	Score / Autograd	$R^3SO(3)$	Vector
Yim *et al.* [68]	$R^3SO(3)$	$\mathcal{N}_{\mathbb{R}^3} \times IG_{SO(3)}$	✗	Score / Autograd	$\langle \mathbb{R}^3, \mathfrak{so}(3) \rangle$	Vector
Ours	$SE(3)$	$\mathcal{N}_{SE(3)}$	$\triangle$	Score / Closed Form	$SE(3)$	Image

groups, distributions, methods, as well as diffusion spaces each method utilizes. Several earlier studies [28, 32] have introduced techniques that operate within the $SO(3)$ space, and adopted normal distributions defined on $SO(3)$ [40] (denoted as $IG_{SO(3)}$). Unfortunately, a primary drawback of $IG_{SO(3)}$ is its absence of a closed form, which poses challenges in its computational efficiency. In a similar vein, the authors in [68] developed a method that operates in the tangent space of $R^3SO(3)$. This method's distribution also does not possess a closed form, which complicates the computational procedure. On the other hand, the authors in [59] employed a joint Gaussian distribution within the $\mathbb{R}^3$ and $SO(3)$ spaces. This distribution benefits from the presence of a closed form and thus offers the potential for increased computational efficiency. However, this approach is confined to the $\mathbb{R}^3 \times SO(3)$ space and treats rotation and translation as separate entities for diffusion. As a result, it may not be able to offer the advantages that $SE(3)$ can provide.

4. Methodology

Given an RGB image I that displays the object of interest, our goal is to estimate the 6D object poses $X = (R, T) \in SE(3)$, which represent the transformation from the camera frame to the object. This estimation involves sampling poses from a conditional distribution $X \sim p(X|I)$, which captures the inherent pose uncertainty of the object depict in I. To facilitate this process, our method employs a score-based generative model on $SE(3)$ to recover this underlying distribution. Poses are then sampled via a *reverse* process that gradually refines noisy pose hypotheses $\tilde{X} \sim p(\tilde{X})$ drawn from a known prior distribution $p(\tilde{X})$, specifically a Gaussian distribution on $SE(3)$. Both the *forward* and *reverse* processes are performed on Lie groups and leverage the associated group operations. It is important to note that our approach does not utilize 3D models of the objects or symmetry annotations during either the training or inference phases, instead relying exclusively on RGB images and the associated ground truth (GT) poses for training.

4.1. Score-Based Pose Diffusion on a Lie Group

To apply score-based generative modeling to a Lie group $\mathcal{G}$, we first establish a perturbation kernel on $\mathcal{G}$ that conforms to the Gaussian distribution [8, 52]. The kernel is given by:

$$p_\Sigma(Y|X) := \mathcal{N}_\mathcal{G}(Y; X, \Sigma) \triangleq \frac{1}{\zeta(\Sigma)} \exp\left(-\frac{1}{2}\mathrm{Log}(X^{-1}Y)^\top \Sigma^{-1} \mathrm{Log}(X^{-1}Y)\right), \quad (3)$$

where Σ is the covariance matrix with diagonal entries populated by σ for representing the scale of the perturbation, $\zeta(\Sigma)$ is the normalizing constant, and $X, Y \in \mathcal{G}$ denote the group elements. The *score* on $\mathcal{G}$ then corresponds to the gradient of the log-density of the data distribution with respect to the group element Y. It can be formulated as follows:

$$\nabla_Y \log p_\Sigma(Y|X) = -\mathbf{J}_r^{-\top}(\mathrm{Log}(X^{-1}Y))\Sigma^{-1}\mathrm{Log}(X^{-1}Y). \quad (4)$$

This term can be expressed in closed form if the inverse of the right-Jacobian $\mathbf{J}_r^{-1}$ on $\mathcal{G}$ exists in a closed form. Nevertheless, an alternative approach suggested by the authors in [59] would be to compute this term using automatic differentiation [43]. By substituting Y with $\tilde{X}$, assuming $\tilde{X} = X\mathrm{Exp}(z),\ z \sim \mathcal{N}(0, \sigma_i^2 I)$, and integrating the above definition, the *score* on $\mathcal{G}$ can be reformulated as follows:

$$\nabla_{\tilde{X}} \log p_\sigma(\tilde{X}|X) = -\frac{1}{\sigma^2}\mathbf{J}_r^{-\top}(z)z. \quad (5)$$

A score model $s_{\boldsymbol{\theta}}(\tilde{X}, \sigma)$ can then be trained using the DSM objective shown in Eq. (1), which takes the following form:

$$\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}; \sigma) \triangleq \frac{1}{2}\mathbb{E}_{p_{\text{data}}(X)}\mathbb{E}_{\tilde{X}\sim\mathcal{N}_\mathcal{G}(X,\Sigma)}\left[\left\| s_{\boldsymbol{\theta}}(\tilde{X}, \sigma) - \nabla_{\tilde{X}} \log p_\sigma(\tilde{X}|X)\right\|_2^2\right]. \quad (6)$$

For the denoising process, we employ a variant of the Geodesic Random Walk [5], tailored to the Lie group context, as a means to generate a sample from a noise distribution. The procedure is expressed as follows:

$$\tilde{X}_{i+1} = \tilde{X}_i \mathrm{Exp}(\epsilon_i s_\theta(\tilde{X}_i, \sigma_i) + \sqrt{2\epsilon_i} z_i), \quad z_i \sim \mathcal{N}(0, I). \quad (7)$$

4.2. Efficient Computation of the Stein Score

Even with the above derivation, obtaining the closed-form *score* remains challenging due to its dependency on the selected distribution. For instance, deriving the closed-form *score* for the $IG_{SO(3)}$ distribution [40] poses difficulties. Furthermore, computing the *score* depends on the existence of a closed-form expression for the Jacobian matrix on $\mathcal{G}$. Even if such an expression exists, it may not guarantee computational efficiency compared to automatic differentiation. Therefore, we next discuss a simplification method of the Stein *score* under certain conditions for reducing computational costs on $\mathcal{G}$. This can be expressed in a closed-form if the Jacobian matrix on $\mathcal{G}$ is invertible and if the left and right Jacobian matrices conform to the following relation:

$$\mathbf{J}_l(z) = \mathbf{J}_r^\top(z), \quad \mathbf{J}_l^{-1}(z) = \mathbf{J}_r^{-\top}(z), \quad (8)$$

where $z \in \mathfrak{g}$. As pointed out by [53], $SO(3)$ exhibits this property. Its closed-form *score* can then be simplified by utilizing the following property, which holds on any $\mathcal{G}$ as $\mathbf{J}_l(z)z = z$. The derivation is in the supplementary material. The *score* on $SO(3)$ can then be expressed as follows:

$$\nabla_{\tilde{X}} \log p_\sigma(\tilde{X}|X) = -\frac{1}{\sigma^2}\mathbf{J}_l^{-1}(z)z = -\frac{1}{\sigma^2}z. \quad (9)$$

This shows that the *score* on $SO(3)$ can be simplified to the sampled Gaussian noise z scaled by $-1/\sigma^2$, thus eliminating the need for both automatic differentiation and Jacobian calculations. Similarly, the *score* on $R^3SO(3)$ also has a closed-form as its Jacobians satisfy the relations in Eq. (8):

$$\mathbf{J}_l(z) = (I, \mathbf{J}_l(\phi)) = (I, \mathbf{J}_r^\top(\phi)) = \mathbf{J}_r^\top(z), \quad (10)$$

where in the case of $R^3SO(3)$, $z = (T, \phi) \in \langle \mathbb{R}^3, \mathfrak{so}(3) \rangle$. This implies that the *score* on $R^3SO(3)$ can also be simplified according to the formulation represented by Eq. (9).

4.3. Surrogate Stein Score Calculation on $SE(3)$

While the *score* on $SO(3)$ and $R^3SO(3)$ can be simplified as described in the preceding sections, it can be shown that $SE(3)$ does not possess the property in Eq. (8). Consider the inverse of the left-Jacobian on $SE(3)$ at $z = (\rho, \phi) \in \mathfrak{se}(3)$, expressed as $\mathbf{J}_l^{-1}(z) = \begin{bmatrix} \mathbf{J}_l^{-1}(\phi) & \mathbf{Z}(\rho,\phi) \\ 0 & \mathbf{J}_l^{-1}(\phi) \end{bmatrix}$, where $\mathbf{Z}(\rho, \phi) = -\mathbf{J}_l^{-1}(\phi)\mathbf{Q}(\rho, \phi)\mathbf{J}_l^{-1}(\phi)$. The complete form of $\mathbf{Q}(\rho, \phi)$ can be found in [4, 53] and our supplementary material. The property $\mathbf{Q}^\top(-\rho, -\phi) = \mathbf{Q}(\rho, \phi)$, as derived in the references, leads to the following inequality:

$$\mathbf{J}_r^{-\top}(z) = (\mathbf{J}_l^{-1}(-z))^\top = \begin{bmatrix} \mathbf{J}_l^{-1}(\phi) & 0 \\ \mathbf{Z}(\rho,\phi) & \mathbf{J}_l^{-1}(\phi) \end{bmatrix} \neq \mathbf{J}_l^{-1}(z). \quad (11)$$

This inequality indicates the potential discrepancy between the *score* vector and the denoising direction due to the curvature of the manifold, which may impede the convergence of the reverse process and necessitate additional denoising steps. To address this problem, we turn to higher-order approximation methods by breaking one step of reverse process into multiple smaller sub-steps. Fig. 2 (right) illustrates this one-step denoising process on $SE(2)$ from a noisy sample $\tilde{X} = X\text{Exp}(z)$ to its cleaned counterpart X, with contour lines representing the distance to X in 2D Euclidean space. We observe that increasing the number of sub-steps eventually leads the integral of those *small* transformations approaches the inverse of z. As a result, we propose substituting the *true score* in Eq. (5) with a *surrogate score* in our training objective of Eq. (6) on $SE(3)$, defined as follows:

$$\tilde{s}_X(\tilde{X}, \sigma) \triangleq -\frac{1}{\sigma^2}z. \quad (12)$$

Note that the detailed training and sampling procedures are described and elaborated in our supplementary material.

4.4. The Proposed Framework

Fig. 2 (left) presents an overview of our framework, which consists of a conditioning part and a denoising part. The conditioning part is responsible for generating the condition variable c, which is crucial for guiding the denoising process. This variable c can be derived either from an image encoder which extracts features from an image, or from a positional embedding module [60] that encodes a time index i. In our experiments, we employ ResNet [14] as the image encoder. The separation of the two parts in our framework eliminates the need of image feature extraction in every denoising step, which offers efficiency in the inference phase. For the denoising part, our score model is composed of multiple multi-layer perceptron (MLP) blocks. This structure is inspired by the recent conditional generative models [16, 55], while we have modified their approaches by substituting linear layers for the convolutional ones. The score model processes a noisy pose $\tilde{x}_i \in \mathfrak{g}$ embedded using a positional encoding. It then computes an estimated *score* $s_\theta(\tilde{x}_i, \sigma_i)$. This estimated *score* is subsequently utilized in the denoising process (i.e., Eq. (7)). Please note that the input and output of the denoising part are represented in vector forms within the corresponding Lie algebra space.

Regarding the design of the conditioning mechanism in MLPs, a few prior studies [16, 55] employ scale-bias condition, which is formulated as $f(x, c) = \mathbf{A}(c)x + \mathbf{B}(c)$. Nevertheless, our empirical observations suggest that this conditioning mechanism does not perform satisfactorily when learning distributions on $SO(3)$. This may be attributable to the limited expressivity of the underlying neural networks. Inspired by [33, 70], we introduce a modified Fourier-based conditioning mechanism, which is formulated as follows:

$$f_i(x, c) = \sum_{j=0}^{d-1} \mathbf{W}_{ij} \left(\mathbf{A}_j(c) \cos(\pi x_j) + \mathbf{B}_j(c) \sin(\pi x_j) \right), \quad (13)$$

where d represents the dimension of our linear layer. This form bears similarity to the Fourier series $f(t) = \sum_{k=0}^{\infty} \mathbf{A}_k \cos\left(\frac{2\pi kt}{P}\right) + \mathbf{B}_k \sin\left(\frac{2\pi kt}{P}\right)$. Our motivation stems from the fact that the pose distribution on SO(3) is circular, and can therefore be represented as periodic functions. By the definition of periodic functions, their derivatives are also periodic. It is worth noting that this conditioning mechanism does not introduce additional parameters in our neural network design, as $\mathbf{W}_{ij}$ is provided by the subsequent linear layer. Our experimental findings suggest that this conditioning scheme enhances the ability of neural network to capture periodic features of score fields on $SO(3)$.

5. Experimental Results

In this section, we demonstrate that our score-based diffusion model can produce precise pose estimation on both

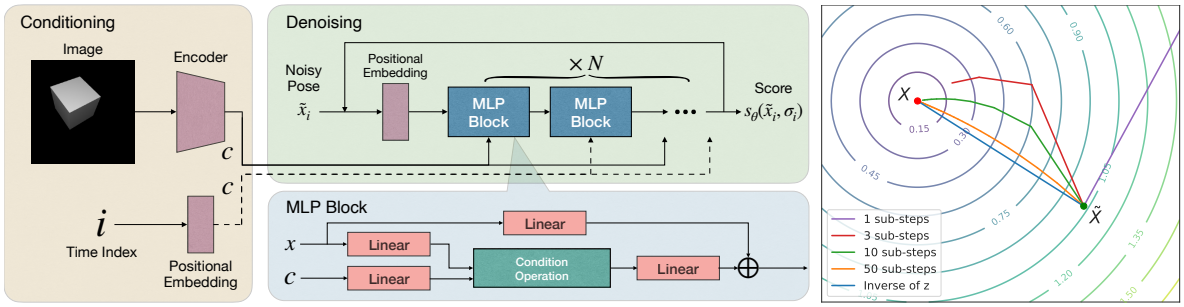

Figure 2. Left: Framework overview. **Right:** Visualization of a denoising step from a noisy sample $\tilde{X}$ to its cleaned counterpart X on $SE(2)$. The contours are the distances to X in 2D Euclidean space. Each line represents a denoising path with varying sub-sampling steps.

$SO(3)$ and $SE(3)$ compared with previous probabilistic approaches. In addition, we present our method's superior performance on the real-world T-LESS [20] dataset without relying on reconstructed 3D models or symmetric annotations. Note that, to the best of our knowledge, our approach is the first probabilistic model that conduct the experiments on the complete T-LESS dataset and reports the accuracy, in contrast to previous methods confined to a limited subset of objects. The extensive evaluation substantiate the robustness and scalability of our score-based diffusion model.

5.1. Experimental Setups

SYMSOL. SYMSOL is a dataset specifically designed for evaluating density estimators in the $SO(3)$ space. This dataset, first introduced by [39], comprises 250k images of five texture-less and symmetric objects, with each subject to random rotations. The objects include tetrahedron (tet.), cube, icosahedron (icosa.), cone, and cylinder (cyl.), with each exhibiting unique symmetries that introduce various degrees of pose ambiguity. For this dataset, our score model is compared in the $SO(3)$ space with several recent works [10, 23, 36, 39]. The baseline models compared with utilize a pre-trained ResNet50 [15] as their backbones. Note that we report the average angular distances in degrees.

SYMSOL-T. To extend our evaluation into the $SE(3)$ space, we developed the SYMSOL-T dataset by incorporating random translations based on SYMSOL, which introduces an additional layer of complexity due to perspective-induced ambiguity. Similar to SYMSOL, it features the same five symmetric shapes and the same number of random samples. For SYMSOL-T, we benchmark our proposed methods against two pose regression methods. These two methods are trained using a symmetry-aware loss, but with different strategies: one directly estimates the pose from an image, while the other employs iterative refinement. We report the average angular distances in degrees for rotation and the average distances for translation.

T-LESS. T-LESS [20] has been recognized as a challenging benchmark in the BOP challenge [22], which consists of thirty texture-less industrial objects. The objects in this dataset are characterized by a range of discrete and continuous symmetries. In this dataset, the pose ambiguities arise not only from the intrinsic object symmetries but also the environmental factors such as occlusion and self-occlusion due to its cluttered settings. The T-LESS dataset features a training set with 50k physically based rendering (PBR) [22] images from synthetic images, and an additional 37k images from real-world scanning. The testing set encompasses 10k real-world scanned images. The evaluation methods employed in our study include three standard metrics from the BOP challenge: Maximum Symmetry-Aware Projection Distance (MSPD), Maximum Symmetry-Aware Surface Distance (MSSD), and Visible Surface Discrepancy (VSD). To reflect the emphasis of our work on symmetry, we further introduced symmetry-aware metrics: R@2, R@5, and R@10, which represent predictions with rotational errors within 2, 5, and 10 degrees, respectively. Similarly, T@2, T@5, and T@10 are estimations with translational errors within 2, 5, and 10 centimeters, respectively.

Visualization To visualize the density predictions, we adopt the strategy employed in [39] to represent the rotation densities generated by our model in the $SO(3)$ space. Specifically, we use the Mollweide projection for visualizing the $SO(3)$ space, with longitude and latitude values representing the yaw and pitch of the object's rotation, respectively. The color in the $SO(3)$ space indicates the roll of the object's rotation. The circles denote sets of equivalent poses, with each dot representing a single sample. For each plot, we generate a total of $1,000$ random samples from our model. For the translation part, we illustrate the rendered results of the estimated poses below their original images.

5.2. Quantitative Results on SYMSOL

In this section, we present the quantitative results evaluated on SYMSOL, and compare our diffusion-based methods with non-parametric ones. We assess the performance

Table 2. Evaluation results on SYMSOL.

Methods	SYMSOL (Spread in degrees ↓)					
	Avg.	**tet.**	**cube**	**icosa.**	**cone**	**cyl.**
DBN [10]	22.44	16.70	40.70	29.50	10.10	15.20
Implicit-PDF [39]	3.96	4.60	4.00	8.40	1.40	1.40
HyperPosePDF [23]	1.94	3.27	2.18	3.24	0.55	0.48
Normalizing Flows [36]	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**

Table 3. Evaluation results on SYMSOL-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

Table 4. Evaluation results on T-LESS (Average of 30 objects).

Methods	T-LESS (Accuracy % ↑)								
	MSPD	**MSSD**	**VSD**	**R@2**	**R@5**	**R@10**	**T@2**	**T@5**	**T@10**
GDRNPP [62]	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

of our score model on $SO(3)$ across various shapes using both ResNet34 and ResNet50 as the backbones, with the results reported in Table 2. Our model demonstrates promising performance, consistently surpassing the contemporary non-parametric baseline models. It is observed that our model, even when based on the less complex ResNet34 backbone, is still able to achieve results that exceed those of the other baselines using the more complex ResNet50 backbone. The average angular errors are consistently below 1 degree across all shape categories. The performance further improves when employing ResNet50, which emphasizes the potential robustness and scalability of using diffusion models for addressing the pose ambiguity problem. However, it is important to observe that our model with ResNet50 exhibits a slightly reduced performance for the cone shape compared to the ResNet34 variant. This discrepancy can be attributed to our practice of training a single model across all shapes, a strategy that parallels those adopted by Implicit-PDF [39] and HyperPosePDF [23]. Such an approach may lead to mutual influences among shapes with diverse pose distributions, and potentially compromise optimal performance for certain shapes. This observation highlights opportunities for future improvements to our model, specifically in enhancing its ability to effectively learn from data spanning various domains. Such endeavors would potentially shed light on the diverse complexities associated with distinct shapes and characteristics.

5.3. Quantitative Results on SYMSOL-T

We report the quantitative results obtained from the SYMSOL-T dataset evaluation, as shown in Table 3. The results reveal that our $SE(3)$ and $R^3SO(3)$ score models outperform the pose regression and iterative regression baselines in terms of estimation accuracy. However, the $R^3SO(3)$ score model encounters difficulty when learning

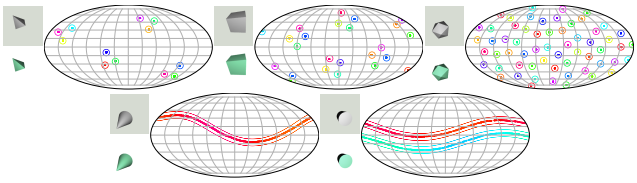

Figure 3. Visualization of our $SE(3)$ diffusion results on SYMSOL-T. Each plot contains $1,000$ sampled poses generated by our model. The first row depicts the densities of discrete symmetrical shapes: (a) tetrahedron, (b) cube, (c) icosahedron, each possessing 12, 24 and 60 discrete symmetries, respectively. The second row presents the densities of continuous symmetrical objects: (d) cone and (e) cylinder, with each shape exhibiting 1 and 2 continuous symmetries, respectively.

Table 5. Inference time (second per sample) across different denoising steps 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

the distribution of the icosahedron shape. In contrast, our $SE(3)$ score model excels in estimating rotation across all shapes and achieves competitive results in translation compared to the $R^3SO(3)$ score model, thus demonstrating its ability to model the joint distribution of rotation and translation. Please note that the $SE(3)$ and $R^3SO(3)$ score models do not rely on symmetry annotations, which distinguish them from the pose regression and iterative regression baselines that leverage symmetry supervision. This supports our initial hypothesis that score models are capable of addressing the pose ambiguity problem in the image domain. In the comparison between the $R^3SO(3)$ score model and iterative regression, both models employ iterative refinement. However, our $R^3SO(3)$ score model consistently outperforms iterative regression on tetrahedron, cube, cone, and cylinder shapes. The key difference is that iterative regression focuses on minimizing pose errors without explicitly learning the underlying true distributions. In contrast, our $R^3SO(3)$ score model captures different scales of noise, enabling it to learn the true distribution of pose uncertainty and achieve more accurate results. Regarding translation performance, the $R^3SO(3)$ score model takes the lead over the $SE(3)$ score model. The former's performance can be credited to its assumption of independence between rotation and translation, which effectively eliminates mutual interference. On the other hand, the $SE(3)$ score model learns the joint distribution of rotation and translation, which leads to more robust rotation estimations. The observations therefore support our hypothesis that $SE(3)$ can provide a more comprehensive pose estimation than $R^3SO(3)$. Fig. 3 show the visualization derived by our model on the $SE(3)$ group.

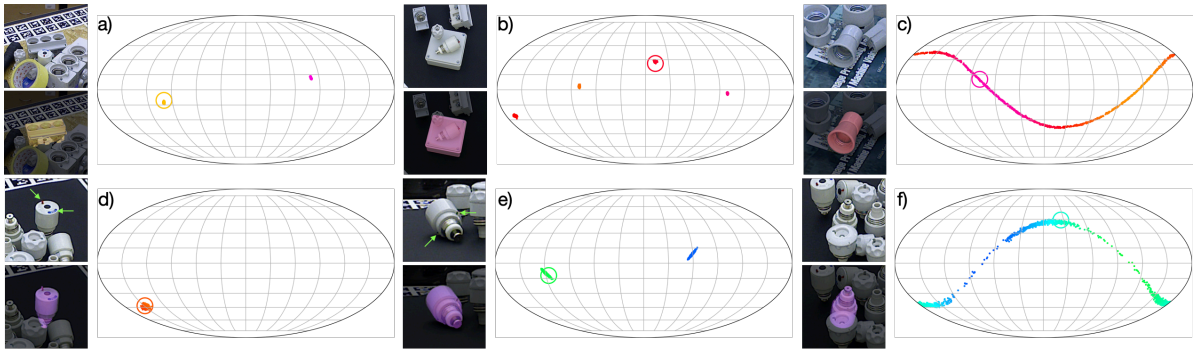

Figure 4. Visualization of our $SE(3)$ diffusion results on T-LESS. In the first row, we present our estimation results of three objects in cluttered scenes: (a) Object 9, characterized by 2 discrete symmetries; (b) Object 27, featuring 4 discrete symmetries; and (c) object 14, possessing 1 continuous symmetries. The second row illustrates pose ambiguities arising from occlusion and self-occlusion, particularly related to Object 4. Notably, this object is annotated with 1 continuous symmetry by human annotator, which does not accurately capture the true ambiguities in certain cases. We explore the scenarios where (d) the object has no symmetry if the top feature is visible; (e) 2 discrete symmetries when the feature is self-occluded, but revealing the two screw holes at the bottom; and (f) 1 continuous symmetry if the screw holes are also occluded by the scene. Each plot contains $1,000$ pose samples from our model. The samples are concentrated on each mode of the distribution, indicating that our models can generate precise rotation estimations across different objects.

5.4. Quantitative Results on T-LESS

We evaluate our $SE(3)$ diffusion model on T-LESS, and demonstrate the effectiveness of our approaches in real-world cluttered scenarios. In this experiment, a single model with a ResNet34 backbone is trained across 30 T-LESS objects. We crop the Region of Interest (RoI) confined within bounding boxes from RGB images and employ segmentation masks to isolate the visible parts of objects. To introduce randomness during training while preserving the RoI aspect ratios, we leverage the Dynamic Zoom-In [35] method. In addition, we apply hard image augmentations [62] to the RoIs, including random colors, Gaussian blur, and noise. It is crucial to note that our method assumes the availability of ground truth bounding boxes and segmentation masks for the visible parts of objects. Table 4 presents the quantitative results. For comparison, we include GDRNPP [62], a regression-based method that stands as the state-of-the-art approach from the BOP challenge in 2022 [57]. The results indicate that our $SE(3)$ diffusion model outperforms its $R^3SO(3)$ counterpart across all metrics. Furthermore, our $SE(3)$ diffusion model demonstrates superior rotation estimation compared to GDRNPP, albeit with a slightly inferior performance in translation. This discrepancy is attributed to GDRNPP's use of geometry guidance derived from 3D models to enhance depth estimation. Fig. 4 presents the visualization results. Please note that more details are presented in the supplementary material.

5.5. Inference Time Analysis

To assess the inference time performance of our models, they are evaluated using the T-LESS dataset and employing JAX [6] as the deep learning package. Our experiments are conducted on an AMD Ryzen Threadripper 2990WX CPU and an RTX 2080 Ti GPU. The models, based on the ResNet34 backbone and an input size of 224 x 224 pixels, demonstrate noticeable efficiency across various denoising steps when parametrized on the $SE(3)$ and $R^3SO(3)$ spaces, as detailed in Table 5. For $SE(3)$, we achieve up to 250 FPS at minimal denoising steps, while for $R^3SO(3)$, the performance reaches 307 FPS. These results suggest the practical applicability of our models in real-time scenarios.

6. Conclusion

In this paper, we presented a novel approach that applies diffusion models to the $SE(3)$ group for object pose estimation, effectively addressing the pose ambiguity issue. Inspired by the correlation between rotation and translation distributions caused by image projection effects, we jointly estimated their distributions on $SE(3)$ for improved accuracy. This is the first work to apply diffusion models to $SE(3)$ in the image domain. To validate it, we developed the SYMSOL-T dataset, which enriches the original SYMSOL dataset with randomly sampled translations. Our experiments confirmed the applicability of our $SE(3)$ diffusion model in the image domain and the advantage of $SE(3)$ parametrization over $R^3SO(3)$. Moreover, our experiments on T-LESS exhibits the efficacy of our $SE(3)$ diffusion model in real-world applications.

7. Acknowledgement

The authors gratefully acknowledge the support from the National Science and Technology Council (NSTC) in Taiwan under grant numbers MOST 111-2223-E-007-004-MY3, Taiwan. The authors would also like to express their appreciation for the donation of the GPUs from NVIDIA Corporation and NVIDIA AI Technology Center (NVAITC) used in this work. Furthermore, the authors extend their gratitude to the National Center for High-Performance Computing (NCHC) for providing the necessary computational and storage resources.

References

[1] Arash Amini, Arul Selvam Periyasamy, and Sven Behnke. Yolopose: Transformer-based multi-object 6d pose estimation using keypoint regression. In *Intelligent Autonomous Systems (IAS)*, pages 392–406, 2022. 3

[2] Tomer Amit, Tal Shaharbany, Eliya Nachmani, and Lior Wolf. Segdiff: Image segmentation with diffusion probabilistic models. *arXiv preprint arXiv:2112.00390*, 2021. 3

[3] Dmitry Baranchuk, Ivan Rubachev, Andrey Voynov, Valentin Khrulkov, and Artem Babenko. Label-efficient semantic segmentation with diffusion models. *arXiv preprint arXiv:2112.03126*, 2021. 3

[4] Timothy D. Barfoot and Paul Timothy Furgale. Associating uncertainty with three-dimensional poses for use in estimation problems. *IEEE Trans. Robotics*, 30:679–693, 2014. 5

[5] Valentin De Bortoli, Emile Mathieu, Michael John Hutchinson, James Thornton, Yee Whye Teh, and Arnaud Doucet. Riemannian score-based generative modelling. In *Advances in Neural Information Processing Systems*, 2022. 3, 4

[6] James Bradbury, Roy Frostig, Peter Hawkins, Matthew James Johnson, Chris Leary, Dougal Maclaurin, George Necula, Adam Paszke, Jake VanderPlas, Skye Wanderman-Milne, and Qiao Zhang. JAX: composable transformations of Python+NumPy programs, 2018. 8

[7] Shoufa Chen, Peize Sun, Yibingimp Song, and Ping Luo. Diffusiondet: Diffusion model for object detection. *CoRR*, abs/2211.09788, 2022. 3

[8] Gregory Chirikjian and Marin Kobilarov. Gaussian approximation of non-linear measurement models on lie groups. In *53rd IEEE Conference on Decision and Control*, pages 6401–6406. IEEE, 2014. 4

[9] Jeongjun Choi, Dongseok Shim, and H. Jin Kim. Diffupose: Monocular 3d human pose estimation via denoising diffusion probabilistic model. *CoRR*, abs/2212.02796, 2022. 3

[10] Haowen Deng, Mai Bui, Nassir Navab, Leonidas Guibas, Slobodan Ilic, and Tolga Birdal. Deep bingham networks: Dealing with uncertainty and ambiguity in pose estimation, 2020. 1, 3, 6, 7

[11] Yan Di, Fabian Manhardt, Gu Wang, Xiangyang Ji, Nassir Navab, and Federico Tombari. So-pose: Exploiting self-occlusion for direct 6d pose estimation. In *Proc. IEEE Int. Conf. on Computer Vision (ICCV)*, pages 12376–12385, 2021. 3

[12] Igor Gilitschenski, Roshni Sahoo, Wilko Schwarting, Alexander Amini, Sertac Karaman, and Daniela Rus. Deep orientation uncertainty learning based on a bingham loss. In *International conference on learning representations*, 2020. 3

[13] Shansan Gong, Mukai Li, Jiangtao Feng, Zhiyong Wu, and LingPeng Kong. Diffuseq: Sequence to sequence text generation with diffusion models. *arXiv preprint arXiv:2210.08933*, 2022. 3

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 5

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 6

[16] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Proc. Conf. on Neural Information Processing Systems (NeurIPS)*, 2020. 1, 2, 3, 5

[17] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022. 3

[18] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *arXiv preprint arXiv:2204.03458*, 2022. 3

[19] Tomáš Hodaň, Pavel Haluza, Štěpán Obdržálek, Jiří Matas, Manolis Lourakis, and Xenophon Zabulis. T-LESS: An RGB-D dataset for 6D pose estimation of texture-less objects. *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2017. 1

[20] Tomáš Hodan, Pavel Haluza, Štepán Obdržálek, Jiri Matas, Manolis Lourakis, and Xenophon Zabulis. T-less: An rgb-d dataset for 6d pose estimation of texture-less objects. In *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 880–888. IEEE, 2017. 2, 6

[21] Tomás Hodan, Dániel Baráth, and Jiri Matas. EPOS: estimating 6d pose of objects with symmetries. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 11700–11709, 2020. 1

[22] Tomáš Hodaň, Martin Sundermeyer, Bertram Drost, Yann Labbé, Eric Brachmann, Frank Michel, Carsten Rother, and Jiří Matas. Bop challenge 2020 on 6d object localization. In *Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 577–594. Springer, 2020. 6

[23] Timon Höfer, Benjamin Kiefer, Martin Messmer, and Andreas Zell. Hyperposepdf-hypernetworks predicting the probability distribution on so (3). In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2369–2379, 2023. 1, 3, 6, 7

[24] Karl Holmquist and Bastian Wandt. Diffpose: Multi-hypothesis human pose estimation using diffusion models. *arXiv preprint arXiv:2211.16487*, 2022. 3

[25] Lin Huang, Tomas Hodan, Lingni Ma, Linguang Zhang, Luan Tran, Christopher D. Twigg, Po-Chen Wu, Junsong Yuan, Cem Keskin, and Robert Wang. Neural correspondence field for object pose estimation. In *Proc. European Conf. on Computer Vision (ECCV)*, pages 585–603, 2022. 3

[26] Rongjie Huang, Zhou Zhao, Huadai Liu, Jinglin Liu, Chenye Cui, and Yi Ren. Prodiff: Progressive fast diffusion model for high-quality text-to-speech. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 2595–2605, 2022. 3

[27] Aapo Hyvärinen and Peter Dayan. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(4), 2005. 2

[28] Yesukhei Jagvaral, Francois Lanusse, and Rachel Mandelbaum. Diffusion generative models on so(3). https:

//openreview.net/pdf?id=jHA-yCyBGb, 2023. 2, 3, 4

[29] Erik Jørgensen. The central limit problem for geodesic random walks. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 32(1-2):1–64, 1975. 3

[30] David M Klee, Ondrej Biza, Robert Platt, and Robin Walters. Image to sphere: Learning equivariant features for efficient pose prediction. *arXiv preprint arXiv:2302.13926*, 2023. 3

[31] Yann Labbé, Justin Carpentier, Mathieu Aubry, and Josef Sivic. Cosypose: Consistent multi-view multi-object 6d pose estimation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVII 16*, pages 574–591. Springer, 2020. 3

[32] Adam Leach, Sebastian M Schmon, Matteo T. Degiacomi, and Chris G. Willcocks. Denoising diffusion probabilistic models on so(3) for rotational alignment. In *Proc. Int. Conf. on Learning Representations Workshop (ICLRW)*, 2022. 2, 3, 4

[33] Jiyoung Lee, Wonjae Kim, Daehoon Gwak, and Edward Choi. Conditional generation of periodic signals with fourier-based decoder. *arXiv preprint arXiv:2110.12365*, 2021. 5

[34] Xiang Li, John Thickstun, Ishaan Gulrajani, Percy S Liang, and Tatsunori B Hashimoto. Diffusion-lm improves controllable text generation. *Advances in Neural Information Processing Systems*, 35:4328–4343, 2022. 3

[35] Zhigang Li, Gu Wang, and Xiangyang Ji. CDPN: coordinates-based disentangled pose network for real-time rgb-based 6-dof object pose estimation. In *Proc. IEEE Int. Conf. on Computer Vision (ICCV)*, pages 7677–7686, 2019. 8

[36] Yulin Liu, Haoran Liu, Yingda Yin, Yang Wang, Baoquan Chen, and He Wang. Delving into discrete normalizing flows on so(3) manifold for probabilistic rotation modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21264–21273, 2023. 6, 7

[37] Yulin Liu, Haoran Liu, Yingda Yin, Yang Wang, Baoquan Chen, and He Wang. Delving into discrete normalizing flows on so (3) manifold for probabilistic rotation modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21264–21273, 2023. 3

[38] Fabian Manhardt, Diego Martín Arroyo, Christian Rupprecht, Benjamin Busam, Tolga Birdal, Nassir Navab, and Federico Tombari. Explaining the ambiguity of object detection and 6d pose from visual data. In *Proc. IEEE Int. Conf. on Computer Vision (ICCV)*, pages 6840–6849, 2019. 1, 3

[39] Kieran A. Murphy, Carlos Esteves, Varun Jampani, Srikumar Ramalingam, and Ameesh Makadia. Implicit-pdf: Nonparametric representation of probability distributionson the rotation manifold. In *Proc. Int. Conf. on Machine Learning (ICML)*, pages 7882–7893, 2021. 1, 2, 3, 6, 7

[40] Dmitry I Nikolayev and Tatjana I Savyolov. Normal distribution on the rotation group so (3). *Textures and Microstructures*, 29, 1970. 4

[41] Brian Okorn, Mengyun Xu, Martial Hebert, and David Held. Learning orientation distributions for object pose estimation. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 10580–10587. IEEE, 2020. 3

[42] Kiru Park, Timothy Patten, and Markus Vincze. Pix2pose: Pixel-wise coordinate regression of objects for 6d pose estimation. In *Proc. IEEE Int. Conf. on Computer Vision (ICCV)*, pages 7667–7676, 2019. 1, 3

[43] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017. 4

[44] Sida Peng, Yuan Liu, Qixing Huang, Xiaowei Zhou, and Hujun Bao. Pvnet: Pixel-wise voting network for 6dof pose estimation. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 4561–4570, 2019. 3

[45] Sergey Prokudin, Peter Gehler, and Sebastian Nowozin. Deep directional statistics: Pose estimation with uncertainty quantification. In *Proceedings of the European conference on computer vision (ECCV)*, pages 534–551, 2018. 3

[46] Mahdi Rad and Vincent Lepetit. BB8: A scalable, accurate, robust to partial occlusion method for predicting the 3d poses of challenging objects without using depth. In *Proc. IEEE Int. Conf. on Computer Vision (ICCV)*, pages 3848–3856, 2017. 3

[47] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 3

[48] Danilo Jimenez Rezende, George Papamakarios, Sébastien Racaniere, Michael Albergo, Gurtej Kanwar, Phiala Shanahan, and Kyle Cranmer. Normalizing flows on tori and spheres. In *International Conference on Machine Learning*, pages 8083–8092. PMLR, 2020. 3

[49] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 3

[50] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. *arXiv preprint arXiv:2208.12242*, 2022.

[51] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022. 3

[52] Salem Said, Lionel Bombrun, Yannick Berthoumieu, and Jonathan H. Manton. Riemannian gaussian distributions on the space of symmetric positive definite matrices. *IEEE Trans. Inf. Theory*, 63(4):2153–2170, 2017. 4

[53] Joan Solà, Jérémie Deray, and Dinesh Atchuthan. A micro lie theory for state estimation in robotics. *CoRR*, abs/1812.01537, 2018. 2, 5

[54] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *Proc. Int. Conf. on Learning Representations (ICLR)*, 2021. 1, 3

[55] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In *Proc. Conf. on Neural Information Processing Systems (NeurIPS)*, pages 11895–11907, 2019. 3, 5

[56] Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *Proc. Int. Conf. on Learning Representations (ICLR)*, 2021. 1, 2, 3

[57] Martin Sundermeyer, Tomáš Hodaň, Yann Labbe, Gu Wang, Eric Brachmann, Bertram Drost, Carsten Rother, and Jiří Matas. Bop challenge 2022 on detection, segmentation and pose estimation of specific rigid objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2784–2793, 2023. 8

[58] Stefan Thalhammer, Timothy Patten, and Markus Vincze. COPE: end-to-end trainable constant runtime object pose estimation. In *Proc. IEEE Winter Conf. on Applications of Computer Vision (WACV)*, pages 2859–2869, 2023. 1, 3

[59] Julen Urain, Niklas Funk, Jan Peters, and Georgia Chalvatzaki. Se(3)-diffusionfields: Learning smooth cost functions for joint grasp and motion optimization through diffusion. *CoRR*, abs/2209.03855, 2022. 2, 3, 4

[60] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 5

[61] Pascal Vincent. A connection between score matching and denoising autoencoders. *Neural Comput.*, 23(7):1661–1674, 2011. 3

[62] Gu Wang, Fabian Manhardt, Federico Tombari, and Xiangyang Ji. Gdr-net: Geometry-guided direct regression network for monocular 6d object pose estimation. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 16611–16621, 2021. 1, 3, 7, 8

[63] He Wang, Srinath Sridhar, Jingwei Huang, Julien Valentin, Shuran Song, and Leonidas J Guibas. Normalized object coordinate space for category-level 6d object pose and size estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2642–2651, 2019. 3

[64] Yu Xiang, Tanner Schmidt, Venkatraman Narayanan, and Dieter Fox. Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes. In *Robotics: Science and Systems XIV*, 2018. 3

[65] Dongchao Yang, Jianwei Yu, Helin Wang, Wen Wang, Chao Weng, Yuexian Zou, and Dong Yu. Diffsound: Discrete diffusion model for text-to-sound generation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2023. 3

[66] Ling Yang, Zhilong Zhang, Yang Song, Shenda Hong, Runsheng Xu, Yue Zhao, Yingxia Shao, Wentao Zhang, Bin Cui, and Ming-Hsuan Yang. Diffusion models: A comprehensive survey of methods and applications. *arXiv preprint arXiv:2209.00796*, 2022. 3

[67] Ruihan Yang, Prakhar Srivastava, and Stephan Mandt. Diffusion probabilistic modeling for video generation. *arXiv preprint arXiv:2203.09481*, 2022. 3

[68] Jason Yim, Brian L. Trippe, Valentin De Bortoli, Emile Mathieu, Arnaud Doucet, Regina Barzilay, and Tommi S. Jaakkola. SE(3) diffusion model with application to protein backbone generation. *CoRR*, abs/2302.02277, 2023. 2, 3, 4

[69] Yingda Yin, Yang Wang, He Wang, and Baoquan Chen. A laplace-inspired distribution on SO(3) for probabilistic rotation estimation. In *The Eleventh International Conference on Learning Representations*, 2023. 3

[70] Liu Ziyin, Tilman Hartwig, and Masahito Ueda. Neural networks fail to learn periodic functions and how to fix it. *Advances in Neural Information Processing Systems*, 33:1583–1594, 2020. 5