Implicit-PDF: Non-Parametric Representation of Probability Distributions on the Rotation Manifold

Kieran Murphy [*] [1] Carlos Esteves [*] [1] Varun Jampani [1] Srikumar Ramalingam [1] Ameesh Makadia [1]

Abstract

Single image pose estimation is a fundamental problem in many vision and robotics tasks, and existing deep learning approaches suffer by not completely modeling and handling: i) uncertainty about the predictions, and ii) symmetric objects with multiple (sometimes infinite) correct poses. To this end, we introduce a method to estimate arbitrary, non-parametric distributions on SO(3). Our key idea is to represent the distributions implicitly, with a neural network that estimates the probability given the input image and a candidate pose. Grid sampling or gradient ascent can be used to find the most likely pose, but it is also possible to evaluate the probability at any pose, enabling reasoning about symmetries and uncertainty. This is the most general way of representing distributions on manifolds, and to showcase the rich expressive power, we introduce a dataset of challenging symmetric and nearly-symmetric objects. We require no supervision on pose uncertainty – the model trains only with a single pose per example. Nonetheless, our implicit model is highly expressive to handle complex distributions over 3D poses, while still obtaining accurate pose estimation on standard non-ambiguous environments, achieving state-of-the-art performance on Pascal3D+ and ModelNet10-SO(3) benchmarks. Code, data, and visualizations may be found at implicit-pdf.github.io.

1. Introduction

There is a growing realization in deep learning that bestowing a network with the ability to express uncertainty is universally beneficial and of crucial importance to systems where safety and interpretability are primary concerns (Leibig et al., 2017; Han et al., 2007; Ching et al., 2018). A quintessential example is the task of 3D pose estimation – pose estimation is both a vital ingredient in many real-world robotics and computer vision applications where propagating uncertainty can facilitate complex downstream reasoning (McAllister et al., 2017), as well as an inherently ambiguous problem due to the abundant approximate and exact symmetries in our 3D world.

Many everyday objects possess symmetries such as the box or vase depicted in Fig. 1 (a). It is tempting to formulate a model of uncertainty that precisely mirrors the pose ambiguities of such shapes; however it becomes immediately evident that such an approach is not scalable, as it is unrealistic to enumerate or characterize all sources of pose uncertainty. Even in a simple scenario such as a coffee mug with self-occlusion, the pose uncertainty manifests as a complex distribution over 3D orientations, as in Fig. 1 (b).

This paper addresses two long-standing and open challenges in pose estimation (a) *what is the most general representation for expressing arbitrary pose distributions, including the challenging ones arising from symmetrical and near-symmetrical objects, in a neural network* and (b) *how do we effectively train the model in typical scenarios where the supervision is a single 3D pose per observation* (as in Pascal3D+ (Xiang et al., 2014), ObjectNet3D (Xiang et al., 2016), ModelNet10-SO(3) (Liao et al., 2019)), i.e. without supervision on the distribution, or priors on the symmetries.

To this end, we propose an *implicit* representation for non-parametric probability distributions over the rotation manifold $\mathbf{SO}(3)$ (we refer to our model as implicit-PDF, or IPDF for short). Such an implicit representation can be parameterized with a neural network and successfully trained with straightforward sampling strategies – uniform or even random querying of the implicit function is sufficient to reconstruct the unnormalized distribution and approximate the normalizing term. For inference, in addition to reconstructing the full probability distribution we can combine the sampling strategy with gradient ascent to make pose predictions at arbitrary (continuous) resolution. The use of a non-parametric distribution, while being simple, offers maximal expressivity for arbitrary densities and poses arising from symmetrical and near symmetrical 3D objects. The simplicity of our approach is in stark contrast to com-

[*]Equal contribution [1]Google Research, New York, NY, USA. Correspondence to: <implicitpdf@gmail.com>.

Proceedings of the 38th International Conference on Machine Learning, PMLR 139, 2021. Copyright 2021 by the author(s).

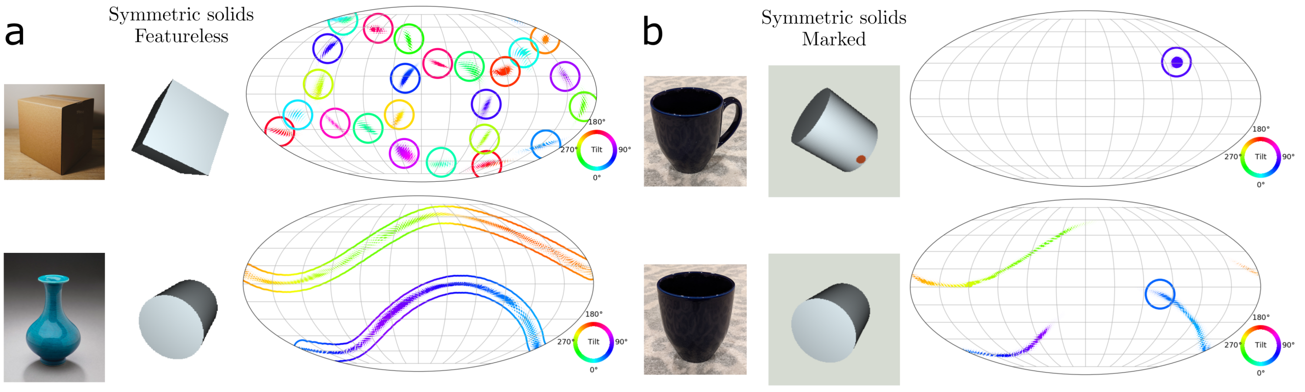

Figure 1. We introduce a method to predict arbitrary distributions over the rotation manifold. This is particularly useful for pose estimation of symmetric and nearly symmetric objects, since output distributions can include both uncertainty on the estimation and the symmetries of the object. ***a-top:*** The cube has 24 symmetries, which are represented by 24 points on $\mathbf{SO}(3)$, and all modes are correctly inferred by our model. ***a-bottom:*** The cylinder has a continuous symmetry around one axis, which traces a cycle on $\mathbf{SO}(3)$. It also has a discrete 2-fold symmetry (a "flip"), so the distribution is represented as two cycles. The true pose distribution for the vase depicted on the left would trace a single cycle on $\mathbf{SO}(3)$ since it does not have a flip symmetry. ***b:*** This cylinder has a mark that uniquely identifies its pose, when visible (top). When the mark is not visible (bottom), our model correctly distributes the probability over poses where the mark is invisible. This example is analogous to a coffee cup when the handle is not visible. The resulting intricate distribution cannot be easily approximated with usual unimodal or mixture distributions on $\mathbf{SO}(3)$, but is easily handled by our IPDF model. *Visualization:* Points with non-negligible probability are displayed as dots on the sphere according to their first canonical axis, colored according to the rotation about that axis. The ground truth (used for evaluation only, not training) is shown as a solid outline. Refer to Section 3.5 for more details.

monly used parametric distributions on $\mathbf{SO}(3)$ that require complicated approximations for computing the normalizing term and further are not flexible enough to fit complex distributions accurately (Gilitschenski et al., 2019; Deng et al., 2020; Mohlin et al., 2020). Our primary contributions are

- *Implicit*-PDF, a novel approach for modeling non-parametric distributions on the rotation manifold. Our implicit representation can be applied to realistic challenging pose estimation problems where uncertainty can arise from approximate or exact symmetries, self-occlusion, and noise. We propose different sampling strategies which allow us to both efficiently reconstruct full distributions on $\mathbf{SO}(3)$ as well as generate multiple pose candidates with continuous precision.
- SYMSOL, a new dataset with inherent ambiguities for analyzing pose estimation with uncertainty. The dataset contains shapes with high order of symmetry, as well as nearly-symmetric shapes, that challenge probabilistic approaches to accurately learn complex pose distributions. When possible, objects are paired with their ground truth "symmetry maps", which allows quantitative evaluation of predicted distributions.

Our IPDF method is extensively evaluated on the new SYMSOL dataset as well as traditional pose estimation benchmarks. To aid our analysis, we develop a novel visualization method for distributions on $\mathbf{SO}(3)$ that provides an intuitive way to qualitatively assess predicted distributions. Through evaluation of predicted distributions and poses, we obtain a broad assessment of our method: IPDF is the only technique that can consistently accurately recover the complex pose uncertainty distributions arising from a high degree of symmetry or self-occlusion, while being supervised by only a single pose per example. Further, while IPDF has the expressive power to model non-trivial distributions, it does not sacrifice in ability to predict poses in non-ambiguous situations and reaches state of the art performance with the usual metrics on many categories of Pascal3D+ (Xiang et al., 2014) and ModelNet10-SO(3) (Liao et al., 2019).

2. Related work

Symmetries are plentiful in our natural and human-made worlds, and so it is not surprising there is a history in computer vision of exploiting strong priors or assumptions on shape or texture symmetry to recover 3D structure from a single image (Poggio & Vetter, 1992; Hong et al., 2004; Rothwell et al., 1993). However, among the more recent machine learning approaches for pose estimation, symmetries are treated as nuisances and strategies have been developed to utilize symmetry annotations at training. With known symmetries at training, a canonical normalization of rotation space unambiguously resolves each set of equivalent rotations to a single one, allowing training to proceed as in single-valued regression (Pitteri et al., 2019). In Corona et al.

(2018), manually annotated symmetries on 3D shapes are required to jointly learn image embedding and classification of the object's symmetry order. Learning representations that cover a few specific symmetry classes is considered in Saxena et al. (2009).

In contrast to these works, Sundermeyer et al. (2019) make pose or symmetry supervision unnecessary by using a denoising autoencoder to isolate pose information. Neither Sundermeyer et al. (2019) nor Corona et al. (2018) directly predict pose, and thus require comparing against many rendered images of the same exact object for pose inference. In a similar vein, Okorn et al. (2020) use a learned comparison against a dictionary of images to construct a histogram over poses. Deng et al. (2019) propose a particle filter framework for 6D object pose tracking, where each particle represents a discrete distribution over $\mathbf{SO}(3)$ with 191K bins. Similar to the previously mentioned works, this discrete rotation likelihood is estimated by codebook matching and an autoencoder is trained to generate the codes.

As noted earlier, symmetries are not the only source of pose uncertainty. Aiming to utilize more flexible representations, a recent direction of work has looked to directional statistics (Mardia & Jupp, 2000) to consider parametric probability distributions. Regression to the parameters of a von Mises distribution over (Euler) angles (Prokudin et al., 2018), as well as regression to the Bingham (Peretroukhin et al., 2020; Deng et al., 2020; Gilitschenski et al., 2019) and Matrix Fisher distributions (Mohlin et al., 2020) over $\mathbf{SO}(3)$ have been proposed. Since it is preferable to train these probabilistic models with a likelihood loss, the distribution's normalizing term must be computed, which is itself a challenge (it is a hypergeometric function of a matrix argument for Bingham and Matrix Fisher distributions). Gilitschenski et al. (2019) and Deng et al. (2020) approximate this function and gradient via interpolation in a lookup table, Mohlin et al. (2020) use a hand-crafted approximation scheme to compute the gradient, and Peretroukhin et al. (2020) simply forgo the likelihood loss. In the simplest setting these models are unimodal, and thus ill equipped to deal with non-trivial distributions. To this end, Prokudin et al. (2018), Gilitschenski et al. (2019), and Deng et al. (2020) propose using multimodal mixture distributions. One challenge to training the mixtures is avoiding mode collapse, for which a winner-take-all strategy can be used (Deng et al., 2020). An alternative to the mixture models is to directly predict multiple pose hypotheses (Manhardt et al., 2019), but this does not share any of the benefits of a probabilistic representation.

Bayesian deep learning provides a general framework to reason about model uncertainty, and in Kendall & Cipolla (2016) test time dropout (Gal & Ghahramani, 2016) was used to approximate Bayesian inference for camera relocalization. Inference with random dropout applied to the trained model is used to generate Monte Carlo pose samples, and thus this approach does not offer a way to estimate the density at arbitrary poses (sampling large numbers of poses would also be impractical).

An alternative framework for representing arbitrary complex distributions is Normalizing Flows (Rezende & Mohamed, 2015). In principle, the reparameterization trick for Lie groups introduced in Falorsi et al. (2019) allows for constructing flows to the Lie algebra of $\mathbf{SO}(3)$. Rezende et al. (2020) develop normalizing flows for compact and connected differentiable manifolds, however it is still unclear how to effectively construct flows on non-Euclidean manifolds, and so far there has been little evidence of a successful application to realistic problems at the complexity of learning arbitrary distributions on $\mathbf{SO}(3)$.

The technical design choices of our implicit pose model are inspired by the very successful implicit shape (Mescheder et al., 2019) and scene (Mildenhall et al., 2020) representations, which can represent detailed geometry with a multilayer perceptron that takes low-dimensional position and/or directions as inputs.

We introduce the details of our approach next.

3. Methods

The method centers upon a multilayer perceptron (MLP) which implicitly represents probability distributions over $\mathbf{SO}(3)$. The input to the MLP is a pair comprising a rotation and a visual representation of an image obtained using a standard feature extractor such as a residual network; the output is an unnormalized log probability. Roughly speaking, we construct the distribution for a given image by populating the space of rotations with such queries, and then normalizing the probabilities. This procedure is highly parallelizable and efficient (see Supp. for time ablations). In the following we provide details for the key ingredients of our method.

3.1. Formalism

Our goal is, given an input $x \in \mathcal{X}$ (for example, an image), to obtain a conditional probability distribution $p(\cdot|x)\colon \mathbf{SO}(3) \mapsto \mathbb{R}^+$, that represents the pose of x. We achieve this by training a neural network to estimate the unnormalized joint log probability function $f\colon \mathcal{X} \times \mathbf{SO}(3) \mapsto \mathbb{R}$. Let α be the normalization term such that $p(x, R) = \alpha \exp(f(x, R))$, where p is the joint distribution. The computation of α is infeasible, requiring integration over $\mathcal{X}$. From the product rule, $p(R|x) = p(x, R)/p(x)$. We estimate $p(x)$ by marginalizing over $\mathbf{SO}(3)$, and since $\mathbf{SO}(3)$ is low-dimensional, we approximate the integral with a discrete sum as follows,

$$\begin{aligned} p(x) &= \int_{R\in\mathbf{SO}(3)} p(x,R)\,dR \\ &= \alpha \int_{R\in\mathbf{SO}(3)} \exp(f(x,R))\,dR \\ &\approx \alpha \sum_i^N \exp(f(x,R_i))V, \end{aligned} \tag{1}$$

where the $\{R_i\}$ are centers of an equivolumetric partitioning of $\mathbf{SO}(3)$ with N partitions of volume $V = {}^{\pi^2}\!/_{N}$. (see Section 3.4 for details). Now α cancels out in the expression for $p(R|x)$, giving

$$p(R|x) \approx \frac{1}{V} \frac{\exp(f(x,R))}{\sum_i^N \exp(f(x,R_i))}, \tag{2}$$

where all the RHS terms are obtained from querying the neural network.

During training, the model receives pairs of inputs x and corresponding ground truth R, and the objective is to maximize $p(R|x)$. See Section 3.3 for details.

Inference – single pose. To make a single pose prediction, we solve

$$R_x^* = \operatorname*{arg\,max}_{R\in\mathbf{SO}(3)} f(x,R), \tag{3}$$

with gradient ascent, since f is differentiable. The initial guess comes from evaluating a grid $\{R_i\}$. Since the domain of this optimization problem is $\mathbf{SO}(3)$, we project the values back into the manifold after each gradient ascent step.

Inference – full distribution. Alternatively, we may want to predict a full probability distribution. In this case $p(R_i|x)$ is evaluated over the $\mathbf{SO}(3)$ equivolumetric partition $\{R_i\}$. This representation allows us to reason about uncertainty and observe complex patterns of symmetries and near-symmetries.

Our method can estimate intricate distributions on the manifold without direct supervision of such distributions. By learning to maximize the likelihood of a single ground truth pose per object over a dataset, with no prior knowledge of each object's symmetries, appropriate patterns expressing symmetries and uncertainty naturally emerge in our model's outputs, as shown in Fig. 1.

3.2. Network

Inspired by recent breakthroughs in implicit shape and scene representations (Mescheder et al., 2019; Park et al., 2019; Sitzmann et al., 2019), we adopt a multilayer perceptron (MLP) to implicitly represent the pose distribution. Differently from most implicit models, we train a single model to represent the pose of any instance of multiple categories, so an input descriptor (e.g. pre-trained CNN features for image inputs) is also fed to the MLP, which we produce with a pre-trained ResNet (He et al., 2015). Most implicit representation methods for shapes and scenes take a position in Euclidean space and/or a viewing direction as inputs. In our case, we take an arbitrary 3D rotation, so we must revisit the longstanding question of how to represent rotations (Levinson et al., 2020). We found it best to use a 3×3 rotation matrix to avoid discontinuities present in other representations (Saxena et al., 2009). Following Mildenhall et al. (2020), we found positionally encoding each element of the input to be beneficial. See the supplement for ablative studies on these design choices.

3.3. Loss

We train our model by minimizing the predicted negative log-likelihood of the (single) ground truth pose. This requires normalizing the output distribution, which we approximate by evaluating Eq. (2) using the method described in Section 3.4 to obtain an equivolumetric grid over $\mathbf{SO}(3)$, in which case the normalization is straightforward. During training, we rotate the grid such that R_0 coincides with the ground truth. Then, we evaluate $p(R_0|x)$ as in Eq. (2), and the loss is simply

$$\mathcal{L}(x,R_0) = -\log(p(R_0|x)) \tag{4}$$

We noticed that the method is robust enough to be trained without an equivolumetric grid; evaluating Eq. (2) for randomly sampled $R_i \in \mathbf{SO}(3)$, provided that one of them coincides with the ground truth, works similarly well. The equivolumetric partition is still required during inference for accurate representation of the probabilities.

3.4. Sampling the rotation manifold

Training and producing an estimate of the most likely pose does not require precise normalization of the probabilities predicted by the network. However, when the distribution is the object of interest (e.g. an accurate distribution will be used in a downstream task), we can normalize by evaluating on a grid of points with equal volume in $\mathbf{SO}(3)$ and approximating the distribution as a histogram.

We employ a method of generating equivolumetric grids developed by Yershova et al. (2010), which uses as its starting point the HEALPix method of generating equal area grids on the 2-sphere (Gorski et al., 2005). A useful property of this sampling is that it is generated hierarchically, permitting multi-resolution sampling if desired.

The Hopf fibration is leveraged to cover $\mathbf{SO}(3)$ by threading a great circle through each point on the surface of a 2-sphere.

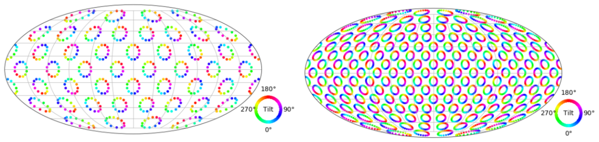

Figure 2. **Equivolumetric grid on SO(3).** In order to normalize the output distribution, we sample unnormalized densities on an equivolumetric grid following Yershova et al. (2010). This iterative method starts with HEALPix (Gorski et al., 2005) which generates equal-area grids hierarchically on the sphere. Left: a grid with 576 samples, right: 4608 samples.

The grids are generated recursively from a starting seed of 72 points, and grow by a factor of eight each iteration. Figure 2 shows grids after one and two subdivisions. For evaluation, we use the grid after 5 subdivisions, with a little more than two million points.

3.5. Visualization

We introduce a novel method to display distributions over $\mathbf{SO}(3)$. A common approach to visualizing such distributions is via multiple marginal distributions, e.g. over each of the three canonical axes (Lee et al., 2008; Mohlin et al., 2020). This is in general incomplete as it is not able to fully specify the joint distribution.

In order to show the full joint distribution, we display the entire space of rotations with the help of the Hopf fibration. With this method, we project a great circle of points on $\mathbf{SO}(3)$ to each point on the 2-sphere, and then use the color wheel to indicate the location on the great circle. More intuitively, we may view each point on the 2-sphere as the direction of a canonical z-axis, and the color indicates the tilt angle about that axis. To represent probability density, we vary the size of the points on the plot. Finally, we display the surface of the 2-sphere using the Mollweide projection.

As the method projects to a lower dimensional space, there are limitations arising from occlusions, but also a freedom in the projection axis which allows finding more or less informative views. The visualization benefits from relatively sparse distributions where much of the space has negligible probability. We did not find this to be limiting in practice: even the 60 modes of a distribution expressing icosahedral symmetry are readily resolved (Fig. 3b).

3.6. Evaluation metrics

The appropriateness of different metrics depends on the nature of predictions (a probability distribution or a set of values) and on the state of knowledge of the ground truth.

Prediction as a distribution: Log likelihood In the most general perspective, ground truth annotations accompanying an image are *observations* from an unknown distribution which incorporates symmetry, ambiguity, and human error involved in the process of annotation. The task of evaluation is a comparison between two distributions given samples from one, for which likelihood is standard (Goodfellow et al., 2014; Clauset et al., 2009; Okorn et al., 2020; Gilitschenski et al., 2019). We report the log likelihood averaged over test set annotations, $\mathbb{E}_{x\sim p(x), R\sim p_{\mathrm{GT}}(R|x)}[\log p(R|x)]$. Importantly, the average log likelihood is invariant to whether one ground truth annotation is available or a set of all equivalent annotations.

Prediction as a distribution: Spread When a complete set of equivalent ground truth values is known (e.g. a value for each equivalent rotation under symmetry), the expected angular deviation to any of the ground truth values is $\mathbb{E}_{R\sim p(R|x)}[\min_{R'\in\{R_{\mathrm{GT}}\}} d(R,R')]$ and $d\colon \mathbf{SO}(3)\times\mathbf{SO}(3)\mapsto\mathbb{R}^+$ is the geodesic distance between rotations. This measure has been referred to as the Mean Absolute Angular Deviation (MAAD) (Prokudin et al., 2018; Gilitschenski et al., 2019), and encapsulates both the deviation from the ground truths and the uncertainty around them.

Prediction as a finite set: precision The most common evaluation scenario in pose estimation tasks is a one-to-one comparison between a single-valued prediction and a ground truth annotation. However, in general, both the prediction and ground truth may be multi-valued, though often only one of the ground truths is available for evaluation. To compensate, sometimes symmetries are implicitly imposed on the entire dataset by reporting flip-invariant metrics (Suwajanakorn et al., 2018; Esteves et al., 2019). These metrics evaluate precision, where a prediction need only be close to one of the ground truths to score well. Usually, the median angular error and accuracy at some angular threshold θ are reported in this setting.

Prediction as a finite set: recall We can also evaluate the coverage of multiple ground truths given multiple predictions, indicating recall. We employ a simple method of clustering by connected components to extract multiple predictions from an output distribution, and rank by probability mass, to return top-k recall metrics; median error and accuracy at θ are evaluated in this setting. When $k = 1$ and the ground truth is unique, these coincide with the precision metrics. See the supplement for extended discussion.

4. Experiments

4.1. Datasets

To highlight the strengths of our method, we put it to the test on a range of challenging pose estimation datasets.

First, we introduce a new dataset (**SYMSOL I**) of images rendered around simple symmetric solids. It includes images of platonic solids (tetrahedron, cube, icosahedron) and

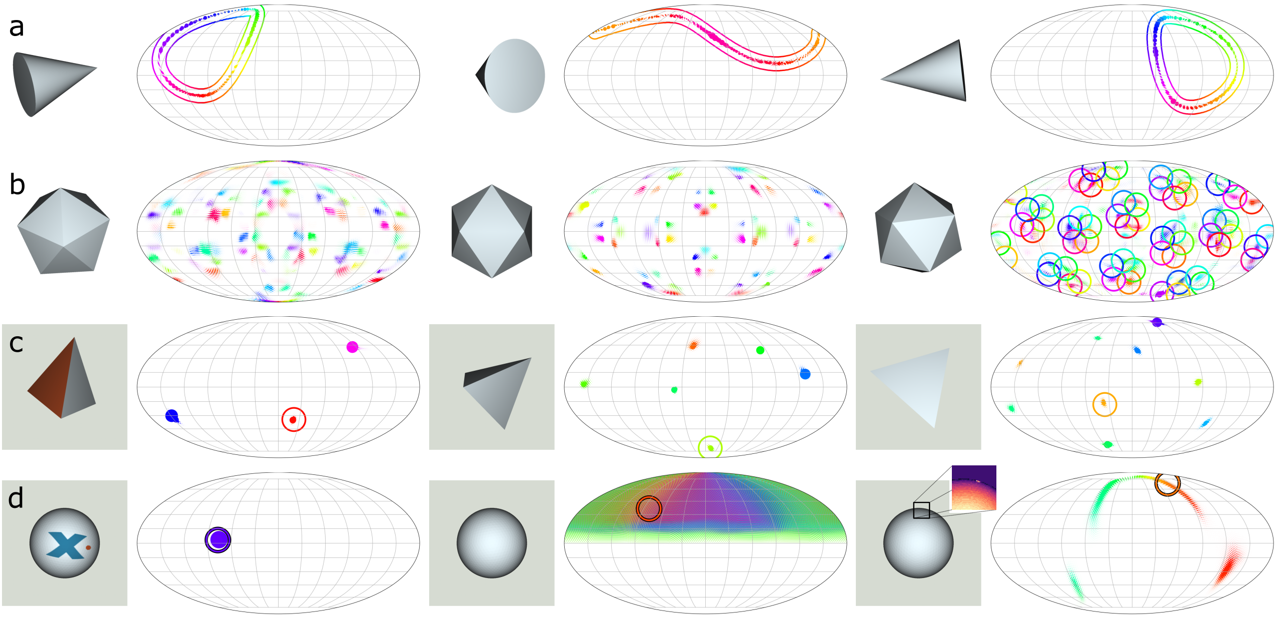

Figure 3. **IPDF predicted distributions for SYMSOL. (a)** The cone has one great circle of equivalent orientations under symmetry. **(b)** The 60 modes of icosahedral symmetry would be exceedingly difficult for a mixture density network based approach, but IPDF can get quite close (we omit the ground truths from the left and middle visualizations for clarity). **(c)** The marked tetrahedron ("tetX") has one red face. When it is visible, the 12 equivalent rotations of a tetrahedron under symmetry reduces to only three. With less information about the location of the red face, more orientations are possible: 6 when two white faces are visible (middle) and 9 when only one white face is visible (right). **(d)** The orientation of the marked sphere ("sphereX") is unambiguous when both markings are visible (left). When they are not (middle), all orientations with the markings on the hidden side of the sphere are possible. When only a portion of the markings are visible (right; inset is a magnification showing several pixels of the X are visible), the IPDF distribution captures the partial information.

surfaces of revolution (cone, cylinder), with 100,000 renderings of each shape from poses sampled uniformly at random from **SO**(3). Each image is paired with its ground truth symmetries (the set of rotations of the source object that would not change the image), which are easily derived for these shapes. As would be the case in most practical situations, where symmetries are not known and/or only approximate, we use such annotations only for evaluation and not for training. Access to the full set of equivalent rotations opens new avenues of evaluating model performance rarely possible with pose estimation datasets.

While the textureless solids generate a challenging variety of distributions, they can still be approximated with mixtures of simple unimodal distributions such as the Bingham (Deng et al., 2020; Gilitschenski et al., 2019). We go one step further and break the symmetry of objects by texturing with small markers (**SYMSOL II**). When the marker is visible, the pose distribution is no longer ambiguous and collapses given the extra information. When the marker is not visible, only a subspace of the symmetric rotations for the textureless shape are possible.

For example, consider a textureless sphere. Its pose distribution is uniform – rotations will not change the input image. Now suppose we mark this sphere with a small arrow. If the arrow is visible, the pose distribution collapses to an impulse. If the arrow is not visible, the distribution is no longer uniform, since about half of the space of possible rotations can now be eliminated. This distribution cannot be easily approximated by mixtures of unimodals.

SYMSOL II objects include a sphere marked with a small letter "X" capped with a dot to break flip symmetry when visible (sphX), a tetrahedron with one red and three white faces (tetX), and a cylinder marked with a small filled off-centered circle (cylO). We render 100,000 images for each.

The two SYMSOL datasets test expressiveness, but the solids are relatively simple and the dataset does not require generalization to unseen objects. **ModelNet10-SO(3)** was introduced by Liao et al. (2019) to study pose estimation on rendered images of CAD models from ModelNet10 (Wu et al., 2015). As in SYMSOL, the rotations of the objects cover all of **SO**(3) and therefore present a difficulty for methods that rely on particular rotation formats such as Euler angles (Liao et al., 2019; Prokudin et al., 2018).

The **Pascal3D+** dataset (Xiang et al., 2014) is a popular benchmark for pose estimation on real images, consisting

Table 1. Distribution estimation on SYMSOL I and II. We report the average log likelihood on both parts of the SYMSOL dataset, as a measure for how well the multiple equivalent ground truth orientations are represented by the output distribution. For reference, a minimally informative uniform distribution over **SO**(3) has an average log likelihood of -2.29. IPDF's expressivity allows it to more accurately represent the complicated pose distributions across all of the shapes. A separate model was trained for each shape for all baselines and for all of SYMSOL II, but only a single IPDF model was trained on all five shapes of SYMSOL I.

	SYMSOL I (log likelihood ↑)						SYMSOL II (log likelihood ↑)			
	avg.	cone	cyl.	tet.	cube	ico.	avg.	sphX	cylO	tetX
Deng et al. (2020)	−1.48	0.16	−0.95	0.27	−4.44	−2.45	2.57	1.12	2.99	3.61
Gilitschenski et al. (2019)	−0.43	3.84	0.88	−2.29	−2.29	−2.29	3.70	3.32	4.88	2.90
Prokudin et al. (2018)	−1.87	−3.34	−1.28	−1.86	−0.50	−2.39	0.48	−4.19	4.16	1.48
IPDF (Ours)	4.10	4.45	4.26	5.70	4.81	1.28	7.57	7.30	6.91	8.49

of twelve categories of objects. Though some of the categories contain instances with symmetries (e.g. bottle and table), the ground truth annotations have generally been disambiguated and restricted to subsets of **SO**(3). This allows methods which regress to a single pose to perform competitively (Liao et al., 2019). Nevertheless, the dataset is a challenging test on real images.

Finally, we evaluate on **T-LESS** (Hodaň et al., 2017), consisting of texture-less industrial parts with various discrete and continuous approximate symmetries. As in Gilitschenski et al. (2019), we use the Kinect RGB single-object images, tight-cropped and color-normalized. Although the objects are nearly symmetric, their symmetry-breaking features are visible in most instances. Nonetheless, it serves as a useful benchmark to compare distribution metrics with Gilitschenski et al. (2019).

We find that IPDF proves competitive across the board.

4.2. Baselines

We compare to several recent works which parameterize distributions on $\mathbf{SO}(3)$ for the purpose of pose estimation. Gilitschenski et al. (2019) and Deng et al. (2020) output the parameters for mixtures of Bingham distributions and interpolate from a large lookup table to compute the normalization constant. Mohlin et al. (2020) output the parameters for a unimodal matrix Fisher distribution and similarly employ an approximation scheme to compute the normalization constant. Prokudin et al. (2018) decompose **SO**(3) into the product of three independent distributions over Euler angles, with the capability for multimodality through an 'infinite mixture' approach. Finally we compare to the spherical regression work of Liao et al. (2019), which directly regresses to Euler angles, to highlight the comparative advantages of distribution-based methods. We quote reported values and run publicly released code when values are unavailable. See Supplemental Material for additional details.

Table 2. ModelNet10-SO(3) accuracy and median angle error. Metrics are averaged over categories. Our model can output pose candidates, so we also evaluate top-k metrics, which are more robust to the lack of symmetry annotations in this dataset. See Supplementary Material for the complete table with per-category metrics.

	Acc@15°↑	Acc@30°↑	Med. (°) ↓
Liao et al. (2019)	0.496	0.658	28.7
Deng et al. (2020)	0.562	0.694	32.6
Prokudin et al. (2018)	0.456	0.528	49.3
Mohlin et al. (2020)	0.693	0.757	17.1
IPDF (ours)	0.719	0.735	21.5
IPDF (ours), top-2	0.868	0.888	4.9
IPDF (ours), top-4	0.904	0.926	4.8

4.3. SYMSOL I: symmetric solids

We report the average log likelihood in Table 1, and the gap between IPDF and the baselines is stark. The average log likelihood indicates how successful the prediction is at distributing probability mass around *all* of the ground truths. The expressivity afforded by our method allows it to capture both the continuous and discrete symmetries present in the dataset. As the order of the symmetry increases from 12 for the tetrahedron, to 24 for the cube, and finally 60 for the icosahedron, the baselines struggle and tend to perform at same level as a minimally informative (uniform) distribution over **SO**(3). The difference between IPDF and the baselines in Table 1 is further cemented by the fact that a single IPDF model was trained on all five shapes while the baselines were allowed a separate model per shape. Interestingly, while the winner-take-all strategy of Deng et al. (2020) enabled training with more Bingham modes than Gilitschenski et al. (2019), it seems to have hindered the ability to faithfully represent the continuous symmetries of the cone and cylinder, as suggested by the relative performance of these methods.

Table 3. Results on a standard pose estimation benchmark, Pascal3D+. As is common, we show accuracy at 30° (top) and median error in degrees (bottom), for each category and also averaged over categories. Our IPDF is at or near state-of-the-art on many categories. ‡ The results for Liao et al. (2019) and Mohlin et al. (2020) differ from their published numbers. For Liao et al. (2019), published errors are known to be incorrectly scaled by a $\sqrt{2}$ factor, and Mohlin et al. (2020) evaluates on a non-standard test set. See Supplemental for details.

		avg.	aero	bike	boat	bottle	bus	car	chair	table	mbike	sofa	train	tv
Acc@30°↑	‡Liao et al. (2019)	0.819	0.82	0.77	0.55	0.93	0.95	0.94	0.85	0.61	0.80	0.95	0.83	0.82
	‡Mohlin et al. (2020)	0.825	0.90	0.85	0.57	0.94	0.95	0.96	0.78	0.62	0.87	0.85	0.77	0.84
	Prokudin et al. (2018)	0.838	0.89	0.83	0.46	0.96	0.93	0.90	0.80	0.76	0.90	0.90	0.82	0.91
	Tulsiani & Malik (2015)	0.808	0.81	0.77	0.59	0.93	0.98	0.89	0.80	0.62	0.88	0.82	0.80	0.80
	Mahendran et al. (2018)	0.859	0.87	0.81	0.64	0.96	0.97	0.95	0.92	0.67	0.85	0.97	0.82	0.88
	IPDF (Ours)	0.837	0.81	0.85	0.56	0.93	0.95	0.94	0.87	0.78	0.85	0.88	0.78	0.86
Median error (°) ↓	‡Liao et al. (2019)	13.0	13.0	16.4	29.1	10.3	4.8	6.8	11.6	12.0	17.1	12.3	8.6	14.3
	‡Mohlin et al. (2020)	11.5	10.1	15.6	24.3	7.8	3.3	5.3	13.5	12.5	12.9	13.8	7.4	11.7
	Prokudin et al. (2018)	12.2	9.7	15.5	45.6	5.4	2.9	4.5	13.1	12.6	11.8	9.1	4.3	12.0
	Tulsiani & Malik (2015)	13.6	13.8	17.7	21.3	12.9	5.8	9.1	14.8	15.2	14.7	13.7	8.7	15.4
	Mahendran et al. (2018)	10.1	8.5	14.8	20.5	7.0	3.1	5.1	9.3	11.3	14.2	10.2	5.6	11.7
	IPDF (Ours)	10.3	10.8	12.9	23.4	8.8	3.4	5.3	10.0	7.3	13.6	9.5	6.4	12.3

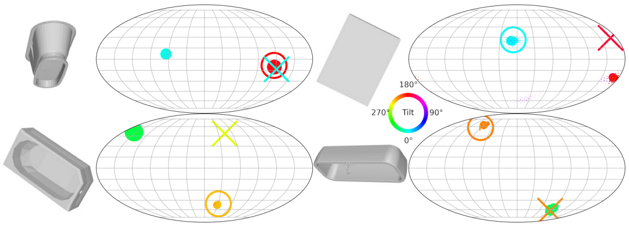

Figure 4. Bathtubs may have exact or approximate 2-fold symmetries around one or more axes. We show our predicted probabilities as solid disks, the ground truth as circles, and the predictions of Liao et al. (2019) as crosses. Our model assigns high probabilities to all symmetries, while the regression method ends up far from every symmetry mode (note the difference in position and color between circles and crosses).

4.4. SYMSOL II: nearly-symmetric solids

When trained on the solids with distinguishing features which are visible only from a subset of orientations, IPDF is far ahead of the baselines (Table 1). The prediction serves as a sort of 'belief state', with the flexibility of being unconstrained by a particular parameterization of the distribution. The marked cylinder in the right half of Figure 1 displays this nicely. When the red marking is visible, the pose is well defined from the image and the network outputs a sharp peak at the correct, unambiguous location. When the cylinder marking is not visible, there is irreducible ambiguity conveyed in the output with half of the full cylindrical symmetry shown in the left side of the figure.

The pose distribution of the marked tetrahedron in Figure 3c takes a discrete form. Depending on which faces are visible, a subset of the full 12 tetrahedral symmetries can be ruled out. For example, with the one red face visible in the left subplot of Figure 3c, there is nothing to distinguish the three remaining faces, and the implicit distribution reflects this state with three modes.

Figure 3d show the IPDF prediction for various views of the marked sphere. When the marking is not visible at all (middle subplot), the half of $\mathbf{SO}(3)$ where the marking faces the camera can be ruled out; IPDF assigns zero probability to half of the space. When only a portion of the marking is visible (right subplot), IPDF yields a nontrivial distribution with an intermediate level of ambiguity, capturing the partial information contained in the image.

4.5. ModelNet10-SO(3)

Unimodal methods perform poorly on categories with rotational symmetries such as *bathtub*, *desk* and *table* (see the supplementary material for complete per-category results). When trained with a single ground truth pose selected randomly from among multiple distinct rotations, these methods tend to split the difference and predict a rotation equidistant from all equivalent possibilities. The most extreme example of this behavior is the *bathtub* category, which contains instances with approximate or exact two-fold symmetry around one or more axes (see Fig. 4). With two modes of symmetry separated by 180°, the outputs tend to be 90° away from each mode. We observe this behavior in Liao et al. (2019); Mohlin et al. (2020).

Since our model can easily represent any kind of symmetry, it does not suffer from this problem, as illustrated in Fig. 4. The predicted distribution captures the symmetry of the object but returns only one of the possibilities during inference. This is penalized by metrics that rely on a single ground truth, since picking the mode that is not annotated

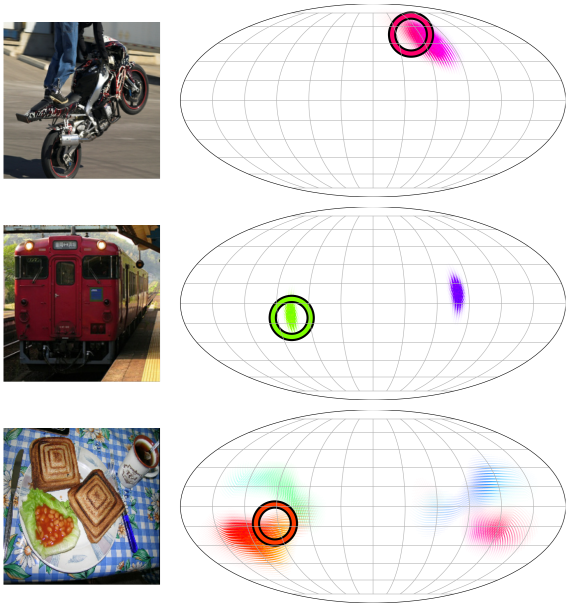

Figure 5. **IPDF predicted distributions on Pascal3D+.** We display a sampling of IPDF pose predictions to highlight the richness of information contained in the full distribution output, as compared to a single pose estimate. Uncertainty regions and multi-modal predictions are freely expressed, owing to the non-parametric nature of IPDF.

results in an 180° error, while picking the midpoint between two modes (which is far from both) results in a 90° error. Since some *bathtub* instances have two-fold symmetries over more than one axis (like the top-right of Fig. 4), our median error ends up closer to 180° when the symmetry annotation is incomplete, which in turn significantly increases the average over all categories. We observe the same for other multi-modal methods (Prokudin et al., 2018; Deng et al., 2020).

Our performance increases dramatically in the top-k evaluation even for $k = 2$ (see Table 2). The ability to output pose candidates is an advantage of our model, and is not possible for direct regression (Liao et al., 2019) or unimodal methods (Mohlin et al., 2020). While models based on mixtures of unimodal distributions could, in theory, produce pose candidates, their current implementations (Gilitschenski et al., 2019; Deng et al., 2020) suffer from mode collapse and are constrained to a fixed number of modes.

4.6. Pascal3D+

In contrast to the full coverage of $\mathbf{SO}(3)$ and the presence of symmetries and ambiguities in the SYMSOL and ModelNet10-SO(3) datasets, Pascal3D+ serves as a check that pose estimation performance in the unambiguous case is not sacrificed. In fact, as the results of Table 3 show, IPDF performs as well as or better than the baselines which constitute a variety of methods to tackle the pose estimation problem. The feat is remarkable given that our method was designed for maximal expressiveness and not for the single-prediction, single-ground truth scenario. IPDF performance in terms of median angular error, while good, overlooks the wealth of information contained in the full predicted distribution. Sample pose predictions are shown in Figure 5 and in the Supplemental; the distributions express uncertainty and category-level pose ambiguities.

Table 4. Pose estimation on T-LESS. LL is the log-likelihood, spread is the mean angular error, and Med. is the median angular error for single-valued predictions. Gilitschenski et al. (2019) underestimate its evaluation of spread, disregarding the dispersion.

	LL ↑	Spread (°) ↓	Med. (°) ↓
Deng et al. (2020)	5.3	23.1	3.1
Gilitschenski et al. (2019)	6.9	3.4	2.7
Prokudin et al. (2018)	8.8	34.3	1.2
Liao et al. (2019)	-	-	2.6
IPDF (Ours)	9.8	4.1	1.3

4.7. T-LESS

The results of Table 4, and specifically the success of the regression method of Liao et al. (2019), show that approximate or exact symmetries are not an issue in the particular split of the T-LESS dataset used in Gilitschenski et al. (2019). All methods are able to achieve median angular errors of less than 4°. Among the methods which predict a probability distribution over pose, IPDF maximizes the average log likelihood and minimizes the spread, when correctly factoring in the uncertainty into the metric evaluation.

5. Conclusion

In this work we have demonstrated the capacity of an implicit function to represent highly expressive, non-parametric distributions on the rotation manifold. It performs as well as or better than state of the art parameterized distribution methods, on standard pose estimation benchmarks where the ground truth is a single pose. On the new and difficult SYMSOL dataset, the implicit method is far superior while being simple to implement as it does not require any onerous calculations of a normalization constant. Particularly, we show in SYMSOL II that our method can represent distributions that cannot be approximated well by current mixture-based models. See the Supplementary Material for additional visualizations, ablation studies and timing evaluations, extended discussion about metrics, and implementation details.

References

Ching, T., Himmelstein, D. S., Beaulieu-Jones, B. K., Kalinin, A. A., Do, B. T., Way, G. P., Ferrero, E., Agapow, P.-M., Zietz, M., Hoffman, M. M., et al. Opportunities and obstacles for deep learning in biology and medicine. *Journal of The Royal Society Interface*, 15 (141):20170387, 2018.

Clauset, A., Shalizi, C. R., and Newman, M. E. Power-law distributions in empirical data. *SIAM review*, 51(4): 661–703, 2009.

Corona, E., Kundu, K., and Fidler, S. Pose Estimation for Objects with Rotational Symmetry. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 7215–7222. IEEE, 2018.

Deng, H., Bui, M., Navab, N., Guibas, L., Ilic, S., and Birdal, T. Deep Bingham Networks: Dealing with Uncertainty and Ambiguity in Pose Estimation. *arXiv preprint arXiv:2012.11002*, 2020.

Deng, X., Mousavian, A., Xiang, Y., Xia, F., Bretl, T., and Fox, D. PoseRBPF: A Rao-Blackwellized Particle Filter for 6D Object Pose Estimation. In *Proceedings of Robotics: Science and Systems*, FreiburgimBreisgau, Germany, June 2019. doi: 10.15607/RSS.2019.XV.049.

Esteves, C., Sud, A., Luo, Z., Daniilidis, K., and Makadia, A. Cross-Domain 3D Equivariant Image Embeddings. In *International Conference on Machine Learning (ICML)*, 2019.

Falorsi, L., de Haan, P., Davidson, T. R., and Forré, P. Reparameterizing Distributions on Lie Groups. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 3244–3253. PMLR, 2019.

Gal, Y. and Ghahramani, Z. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. In *International Conference on Machine Learning (ICML)*, pp. 1050–1059. PMLR, 2016.

Gilitschenski, I., Sahoo, R., Schwarting, W., Amini, A., Karaman, S., and Rus, D. Deep Orientation Uncertainty Learning Based on a Bingham Loss. In *International Conference on Learning Representations*, 2019.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. In *Advances in Neural Information Processing Systems 27*, pp. 2672–2680, 2014.

Gorski, K. M., Hivon, E., Banday, A. J., Wandelt, B. D., Hansen, F. K., Reinecke, M., and Bartelmann, M. HEALPix: A Framework for High-Resolution Discretization and Fast Analysis of Data Distributed on the Sphere. *The Astrophysical Journal*, 622(2):759–771, Apr 2005. ISSN 1538-4357. doi: 10.1086/427976. URL `http://dx.doi.org/10.1086/427976`.

Han, D., Kwong, T., and Li, S. Uncertainties in real-time flood forecasting with neural networks. *Hydrological Processes: An International Journal*, 21(2):223–228, 2007.

He, K., Zhang, X., Ren, S., and Sun, J. Deep Residual Learning for Image Recognition. *arXiv preprint arXiv:1512.03385*, 2015.

Hodaň, T., Haluza, P., Obdržálek, Š., Matas, J., Lourakis, M., and Zabulis, X. T-LESS: An RGB-D Dataset for 6D Pose Estimation of Texture-less Objects. *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2017.

Hong, W., Yang, A. Y., Huang, K., and Ma, Y. On Symmetry and Multiple-View Geometry: Structure, Pose, and Calibration from a Single Image. *International Journal of Computer Vision*, 60(3):241–265, 2004.

Kendall, A. and Cipolla, R. Modelling uncertainty in deep learning for camera relocalization. In *2016 IEEE international conference on Robotics and Automation (ICRA)*, pp. 4762–4769. IEEE, 2016.

Lee, T., Leok, M., and McClamroch, N. H. Global symplectic uncertainty propagation on SO(3). In *Proceedings of the 47th IEEE Conference on Decision and Control, CDC 2008, December 9-11, 2008, Cancún, Mexico*, pp. 61–66, 2008. doi: 10.1109/CDC.2008.4739058. URL `https://doi.org/10.1109/CDC.2008.4739058`.

Leibig, C., Allken, V., Ayhan, M. S., Berens, P., and Wahl, S. Leveraging uncertainty information from deep neural networks for disease detection. *Scientific reports*, 7(1): 1–14, 2017.

Levinson, J., Esteves, C., Chen, K., Snavely, N., Kanazawa, A., Rostamizadeh, A., and Makadia, A. An Analysis of SVD for Deep Rotation Estimation. In *Advances in Neural Information Processing Systems 34*, 2020.

Liao, S., Gavves, E., and Snoek, C. G. M. Spherical Regression: Learning Viewpoints, Surface Normals and 3D Rotations on *n*-Spheres. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

Mahendran, S., Ali, H., and Vidal, R. A mixed classification-regression framework for 3d pose estimation from 2d images. *The British Machine Vision Conference (BMVC)*, 2018.

Manhardt, F., Arroyo, D. M., Rupprecht, C., Busam, B., Birdal, T., Navab, N., and Tombari, F. Explaining the Ambiguity of Object Detection and 6D Pose From Visual

Data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

Mardia, K. V. and Jupp, P. E. *Directional Statistics*. John Wiley and Sons, LTD, London, 2000.

McAllister, R., Gal, Y., Kendall, A., Van Der Wilk, M., Shah, A., Cipolla, R., and Weller, A. Concrete problems for autonomous vehicle safety: Advantages of bayesian deep learning. International Joint Conferences on Artificial Intelligence, Inc., 2017.

Mescheder, L., Oechsle, M., Niemeyer, M., Nowozin, S., and Geiger, A. Occupancy Networks: Learning 3D Reconstruction in Function Space. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, 2019.

Mildenhall, B., Srinivasan, P. P., Tancik, M., Barron, J. T., Ramamoorthi, R., and Ng, R. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. In *ECCV*, 2020.

Mohlin, D., Bianchi, G., and Sullivan, J. Probabilistic Orientation Estimation with Matrix Fisher Distributions. In *Advances in Neural Information Processing Systems 33*, 2020.

Okorn, B., Xu, M., Hebert, M., and Held, D. Learning Orientation Distributions for Object Pose Estimation. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2020.

Park, J. J., Florence, P., Straub, J., Newcombe, R. A., and Lovegrove, S. Deepsdf: Learning continuous signed distance functions for shape representation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pp. 165–174, 2019. doi: 10.1109/CVPR.2019.00025.

Peretroukhin, V., Giamou, M., Rosen, D. M., Greene, W. N., Roy, N., and Kelly, J. A Smooth Representation of SO(3) for Deep Rotation Learning with Uncertainty. In *Proceedings of Robotics: Science and Systems (RSS)*, Jul. 12–16 2020.

Pitteri, G., Ramamonjisoa, M., Ilic, S., and Lepetit, V. On Object Symmetries and 6D Pose Estimation from Images. *CoRR*, abs/1908.07640, 2019. URL http://arxiv.org/abs/1908.07640.

Poggio, T. and Vetter, T. Recognition and Structure from one 2D Model View: Observations on Prototypes, Object Classes and Symmetries. Technical report, MASSACHUSETTS INST OF TECH CAMBRIDGE ARTIFICIAL INTELLIGENCE LAB, 1992.

Prokudin, S., Gehler, P., and Nowozin, S. Deep Directional Statistics: Pose Estimation with Uncertainty Quantification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 534–551, 2018.

Rezende, D. and Mohamed, S. Variational Inference with Normalizing Flows. In *International Conference on Machine Learning*, pp. 1530–1538. PMLR, 2015.

Rezende, D. J., Papamakarios, G., Racaniere, S., Albergo, M., Kanwar, G., Shanahan, P., and Cranmer, K. Normalizing Flows on Tori and Spheres. In *International Conference on Machine Learning*, pp. 8083–8092. PMLR, 2020.

Rothwell, C., Forsyth, D. A., Zisserman, A., and Mundy, J. L. Extracting Projective Structure from Single Perspective Views of 3D Point Sets. In *1993 (4th) International Conference on Computer Vision*, pp. 573–582. IEEE, 1993.

Saxena, A., Driemeyer, J., and Ng, A. Y. Learning 3-D Object Orientation from Images. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2009.

Sitzmann, V., Zollhöfer, M., and Wetzstein, G. Scene representation networks: Continuous 3d-structure-aware neural scene representations. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 1119–1130, 2019.

Sundermeyer, M., Marton, Z., Durner, M., Brucker, M., and Triebel, R. Implicit 3D Orientation Learning for 6D Object Detection from RGB Images. *CoRR*, abs/1902.01275, 2019.

Suwajanakorn, S., Snavely, N., Tompson, J. J., and Norouzi, M. Discovery of Latent 3D Keypoints via End-to-end Geometric Reasoning. In *Advances in Neural Information Processing Systems (NIPS)*, pp. 2063–2074, 2018.

Tulsiani, S. and Malik, J. Viewpoints and keypoints. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

Wu, Z., Song, S., Khosla, A., Yu, F., Zhang, L., Tang, X., and Xiao, J. 3D ShapeNets: A Deep Representation for Volumetric Shapes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1912–1920, 2015.

Xiang, Y., Mottaghi, R., and Savarese, S. Beyond PASCAL: A benchmark for 3D object detection in the wild. In *2014 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 75–82, March 2014.

Xiang, Y., Kim, W., Chen, W., Ji, J., Choy, C., Su, H., Mottaghi, R., Guibas, L., and Savarese, S. ObjectNet3D: A Large Scale Database for 3D Object Recognition. In *European Conference Computer Vision (ECCV)*, 2016.

Yershova, A., Jain, S., Lavalle, S. M., and Mitchell, J. C. Generating Uniform Incremental Grids on SO (3) Using the Hopf Fibration. *The International journal of robotics research*, 29(7):801–812, 2010.