2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 25-29, 2020, Las Vegas, NV, USA (Virtual)

Learning Orientation Distributions for Object Pose Estimation

Brian Okorn[1], Mengyun Xu[1], Martial Hebert[1], David Held[1]

***Abstract*— For robots to operate robustly in the real world, they should be aware of their uncertainty. However, most methods for object pose estimation return a single point estimate of the object's pose. In this work, we propose two learned methods for estimating a distribution over an object's orientation. Our methods take into account both the inaccuracies in the pose estimation as well as the object symmetries. Our first method, which regresses from deep learned features to an isotropic Bingham distribution, gives the best performance for orientation distribution estimation for non-symmetric objects. Our second method learns to compare deep features and generates a non-parameteric histogram distribution. This method gives the best performance on objects with unknown symmetries, accurately modeling both symmetric and non-symmetric objects, without any requirement of symmetry annotation. We show that both of these methods can be used to augment an existing pose estimator. Our evaluation compares our methods to a large number of baseline approaches for uncertainty estimation across a variety of different types of objects. Code available at https://bokorn.github.io/orientation-distributions/**

I. Introduction

Pose estimation is a commonly used primitive in many robotic tasks such as grasping [1], motion planning [2], and object manipulation [3]. For grasping, pose estimation is regularly used to register an observed object to a 3D model for which grasp positions have been annotated [4], [5]. In motion planning, many algorithms require the poses of objects in the environment, either for avoiding collisions [6] or as a state representation used for planning how to manipulate the objects [2].

Most prior methods for pose estimation output a single best guess of each object's pose [7], [8], [9], [10]. In contrast, for many robotic applications, we believe that it is important for a robot to be aware of the uncertainty underlying these estimates before taking an action. This uncertainty can be caused by environmental factors, such as occlusions, poor lighting, or object symmetry, or by biases in the algorithm, induced by insufficient training sets. These factors can cause ambiguity with respect to the object's orientation. If this uncertainty is not taken into account, then the actions of the robot may cause irreversible damage to itself or its environment. For example, a poorly estimated pose estimate can cause a robot to knock over fragile objects while attempting to grasp them. In such cases, rather than taking potentially dangerous actions, the robot should instead capture more information about the environment in an attempt to reduce this uncertainty. Additionally, estimates of uncertainty allow the robot to fuse multiple estimates, through tracking, to achieve a more robust final pose estimate. Thus, methods for pose estimation for robotics should output a distribution of poses rather than just a single pose estimate.

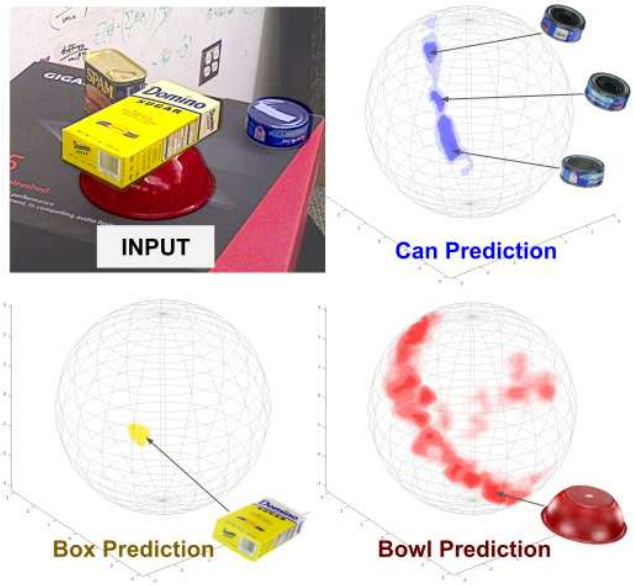

Fig. 1. Multi-modal distributions estimated by our Learned Comparison Histogram approach. These distributions are generated for the tuna can, bowl and sugar box using PoseCNN featurizations of the top right image. Here we see the estimator capturing multiple possible viewpoint for the tuna can, while still placing most of the probability density on the correct mode. It is also able to capture the full symmetry of the bowl without any symmetry labeling. In the case of unambiguous poses, like the sugar box, it is still capable of producing tight uni-modal distributions.

We propose two novel methods for estimating orientation distributions. The first method learns a uni-modal, parametric distribution in the form of an isotropic Bingham, regressed from deep learned features. This model is ideal for objects that are known to be non-symmetric. The second learns to estimate a multi-modal non-parametric distribution, in the form of a histogram distribution, obtained using a learned comparison function over deep learned features. We find that this second method works well for objects with unknown symmetries, accurately modeling both symmetric and non-symmetric objects, without any requirement of symmetry annotation.

We compare our learned methods against other statistically driven methods for estimating parametric and non-parametric orientation distributions. We test each method on the pre-trained feature representations from two state-of-the-art pose estimation methods [7], [8], and evaluate on a large pose estimation dataset [7] that has been used in a number of

*This work was supported by NASA NSTRF, United States Air Force and DARPA under Contract No. FA8750-18-C-0092, National Science Foundation under Grant No. IIS-1849154, and LG Electronics

[1]Brian Okorn, Mengyun Xu, David Held and Martial Hebert are with Robotics Institute at Carnegie Mellon University, 5000 Forbes Ave, Pittsburgh, PA USA; bokorn@andrew.cmu.edu

recent works [8], [11].

II. Related Work

A. Pose Estimation

Previous methods for pose estimation fall into four major categories: segmentation based methods, local coordinate based methods, image template based methods, and direct regression methods. Segmentation based algorithms [12], [13] use an object segmentation algorithm to isolate the points associated with the target object. The segmented depth pixels can be registered with a 3D model of the object using Iterative Closest Point (ICP) algorithms. Local coordinate methods densely predict the 3D location of each pixel with respect to the original object model [9]. These local coordinates define correspondences between the model and the image pixel locations; which are then used with RANSAC [14] to find the object's pose. Alternatively, instead of densely estimating coordinates, the coordinates of an object's bounding box can be regressed can be regressed [11]. Image template methods [15], [16], [17] render a template image at multiple viewpoints around the object model and compute a feature representation at each pose. The objects pose is estimated by looking up the nearest object templates, either by successive pruning of candidates [15], a hashing function [17], [18], or by GPU parallelized comparison [16]. These coarse estimates tend to be refined using ICP. Recently, deep learned methods have been explored, which can directly regress the object's pose using RGB images [7] or densely fused image and point features [8]. Additionally, learned latent spaces have been explored as object pose representations [19], [20], [21]. In this work, we focus not on improving the accuracy of the underlying pose estimate but in adding a model of the estimates uncertainty over the entire orientation space.

B. Pose Distribution Estimation

While most prior methods for pose estimation output a single best guess of each object's pose, there has been some recent work on estimating pose distributions. Su [22] estimated uncertainty distributions over the individual camera view angles relative to classes of objects through a soft classification method. Marton [23] estimated a conditional probability distribution over orientations, in the form of a confusion matrix generated over rendered point clouds. Glover [24] fit mixtures of Bingham distributions to clusters of local point cloud features to estimate an orientation distribution. Similarly, Riedel [25] combined multiple pose estimates using Bingham mixture models. However, unlike this work, they do not evaluate uncertainty estimation with respect to existing deep learned methods or with respect to log likelihood.

Other previous work has estimated a distribution over the object coordinates [26] or bounding box coordinates [11]. However, these methods do not output a distribution over poses, nor do they evaluate whether the distributions themselves are reasonable. One previous paper evaluates distributions over the poses of object classes [22], mostly focusing on azimuth estimation. In contrast, we estimate the orientation distribution of specific object instances and over the full space of orientation.

Most recently, Deng [27] used a learned feature space to estimate multimodal uncertainty distributions over rotations, and used those estimates for particle filter tracking. However, this work did not quantitatively evaluate the uncertainty distribution itself, nor did it compare to other approaches for estimating orientation distributions. Additionally, this method requires the use of a specifically learned autoencoder representation [19]. Manhardt [28] explored learning orientation distributions through PCA analysis of multiple orientation hypotheses, trained using a winner-take-all approach. While this method does visualize their distributions as Bingham distributions, they do not investigate the accuracy of the underlying uncertainty distribution beyond qualitative analysis.

C. Neural Network Uncertainty Estimation

Because deep learning is a popular method for many computer vision tasks (including pose estimation), many approaches have explored how to estimate uncertainty from neural networks. The most popular approaches include Monte Carlo Dropout [29] to estimate epistemic uncertainty, and regressing to the parameters of a distribution [30] to estimate aleatoric uncertainty. We evaluate both of these approaches in this work.

D. Pose Tracking

Tracking 6D rotation has been done using Kalman filters over Bingham Distributions [31], [32]. Bingham distributions [33] are well suited for this problem when the orientation distribution is expected to be unimodal, as they well model rotation quaternion and their composition is well defined. Additionally, particle filtering [27], [34] as well as histogram filtering [23] have been used to sequentially improve and track object pose. The distribution estimates estimated by our method can be similarly used to improve pose estimate accuracy.

III. Background

A. Orientation Representation

Unit quaternions are used as our rotation representation, as they are a compact, numerically stable representation that does not suffer from singularities or gimbal lock. For these reasons, they are the preferred representation of 3D orientation in many papers for both robotics and deep learning [7], [8]. Additionally, unit quaternions have well studied parametric distributions, as well as several uniform sampling strategies [35], [36], [37]. For more background on quaternions, we refer the reader to [38].

B. Bingham distributions

One of our proposed methods, described in Section IV-A, makes use of a Bingham distribution [33]. A Bingham distribution is an antipodal distribution over the surface of a sphere, equivalent to a Gaussian distribution conditioned to lie on the orientation space, SO(3). Bingham distributions

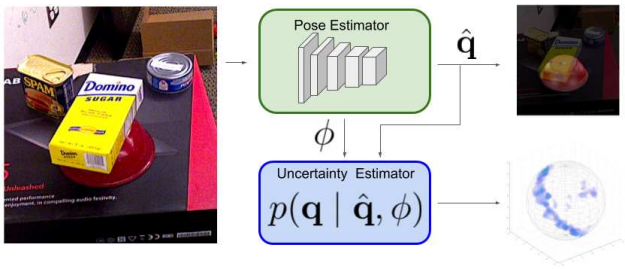

Fig. 2. System pipeline for estimating orientation distributions about an existing pose estimator. The base pose estimator generates an orientation $\hat{\mathbf{q}}$ and a featurization ϕ of the input, one or both of which are used to estimate a uncertainty distribution over possible poses. We render this distribution in as a heat map in axis angle space, lower right, with each orientation being plotted as point in the directions of the axis of rotation and at a distance away form the origin equal to the angle of rotation.

have been used for both orientation tracking and filtering [31], [24], [25]. These distributions are parameterized by an orthogonal 4x4 quaternion rotation matrix $\mathbf{M}$, which describes how the distribution will be rotated on the 3-sphere, and the diagonal 4x4 concentration matrix $\mathbf{Z}$ which describes the spread of the distribution. Similar to Gaussian distributions, Bingham distributions can be simplified to an isotropic distribution, parameterized by a mean quaternion and a single concentration parameter, analogous to variance for a Gaussian) .

IV. Methods for Estimating Orientation Distributions

We introduce two novel algorithms for learning orientation distributions. These methods can be used to augment many existing pose estimators, without decreasing the single point accuracy of the underlying system. In this work, we focus on estimating only the uncertainty of the object's orientation, and not its full 6D pose. However, given a distribution over the object's orientation, a distribution over translation can also be estimated using Rao-Blackwellized particle filter sampling [27].

A. *Bingham Distribution Regression*

Our first method is designed to estimate the distribution of non-symmetric objects. For such objects, we regress the parameters of a Bingham distribution from deep learned object features. Our method builds off of a base pose estimator which extracts a set of features $\phi(I)$ from a cropped image I of the target object. The base pose estimator then regresses from these features $\phi(I)$ to a single point estimate $\bar{\mathbf{q}}$ of the object's orientation. The focus of our approach is not in obtaining these features $\phi(I)$ or in learning the point estimate $\bar{\mathbf{q}}$; rather, these are provided as an input to our system. We evaluate a couple of different options for feature extraction, as explained in Section V-C, and show that our method works for both.

We use the orientation $\bar{\mathbf{q}}$ as the mean of the Bingham distribution. From the features $\phi(I)$, our method learns to regress the remaining parameters of the Bingham distribution, explained below. The parameters of this method are learned by maximizing the log likelihood of the ground-truth pose for each image in the training set.

For simplicity, we limit our Bingham distribution to having an isotropic covariance, requiring only a single parameter σ to be learned. The orthogonality constraint on M can be handled using the Cayleys factorization of the of 4D rotations [39], giving us a parameterization of $\mathbf{M}$ into two unit norm quaternions, $\mathbf{q}_L$ and $\mathbf{q}_R$. By setting $\mathbf{q}_L = \bar{\mathbf{q}}$ and $\mathbf{q}_R$ to the identity quaternion, we both simplify the regression and guarantee that the distribution is centered about $\bar{\mathbf{q}}$. This parameterization can be used to regress an anisotropic Bingham, but we found that the isotropic Bingham produced more accurate results and a more stable training procedure. Results using the full Bingham regression are included as a baseline; see Section V-A.5 for details.

B. *Multi-modal Distribution Regression*

For symmetric objects, or objects that appear symmetric from certain poses or under particular occlusion patterns, a uni-modal Bingham distribution may not be sufficient to capture the object's uncertainty. In such cases, a multi-modal histogram distribution may be more appropriate.

We use a k-nearest neighbor representation over a uniformly gridded space of unique orientations. In this work, we using the discretization method described by Straub [40], as it enforces a near uniform distance between vertices, but any uniform sampling or gridding method could be used. The likelihood estimates at these vertices are interpolated using inverse distance weighting to the k nearest orientations with respect to angular distance. These interpolated values are normalized by dividing by the surface integral of the interpolation over the space of unique rotations, to form a valid continuous probability distribution.

A naive approach to obtaining such a histogram would be to regress from some latent features $\phi(I)$ directly to the parameters of a multi-modal histogram, $p(\mathbf{q} \mid \phi)$. We include

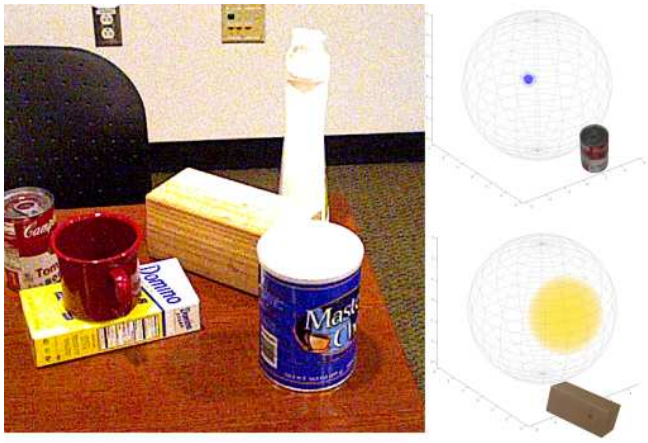

Fig. 3. Isotropic Bingham distributions regressed for the soup can, top, and the wood block, bottom, using DenseFusion featurization. The estimator is able to tightly fit a Bingham to the unambiguous pose of the soup can, but is not able to capture the multi-modal symmetry of the wood block. The only recourse is to inflate the uncertainty in an attempt to capture multiple modes.

this approach as one of our baselines; see Section V-A.6 for details. We show that such a method leads to poor results, due to the inability of such a method to generalize to unseen object viewpoints.

Learned Comparison Histogram: We instead learn a comparison function $f(\phi(I_j) \mid \phi(I))$ between the features $\phi(I)$ and the features $\phi(I_j)$, which are computed from an image of the object rendered at orientation $\mathbf{q}_j$. To simplify notation, we will write this comparison function as $f(\phi_j \mid \phi)$ These rendered orientations are selected using the gridding described above. Our feature comparison function, once normalized, is specifically trained to approximate the posterior, e.g. $f(\phi_j, \phi) \approx \hat{p}(\mathbf{q}_j \mid \phi)$, as described below.

To mimic the posterior $\hat{p}(\mathbf{q}_j \mid \phi)$, we train the comparison function, $f(\phi_j \mid \phi)$, using an interpolated negative log likelihood loss. Specifically, given a ground-truth orientation of $\mathbf{q}^*$, we minimize the loss

$$\mathscr{L}(\mathbf{q}^*; \phi) = -\log\left(\sum_{k=1}^{K} \hat{p}(\mathbf{q}_k \mid \phi)/d(\mathbf{q}^*, \mathbf{q}_k)\right) + \log\left(\sum_{j=1}^{N} \hat{p}(\mathbf{q}_j \mid \phi)\right) \tag{1}$$

where $d(\mathbf{q}^*, \mathbf{q}_k)$ is the minimum angular distance between orientations $\mathbf{q}^*$ and $\mathbf{q}_k$. The set $\{\mathbf{q}_1, \ldots, \mathbf{q}_K\}$ are the K nearest gridded orientations to $\mathbf{q}^*$, and $\{\mathbf{q}_1, \ldots, \mathbf{q}_N\}$ are all of the orientations in our gridding. In our experiments, we use $K = 4$.

We pre-compute the features ϕ_j using a rendered image, I_j, of the object generated with uniform lighting and no occlusions at orientation $\mathbf{q}_j$. This image is then passed through the base pose estimator to extract features ϕ_j. Note that, if the featurization $\phi(\cdot)$ is fixed, the features ϕ_j can be pre-computed and cached. This method is capable of learning tight uni-modal distributions when the pose of the object is unambiguous, like the sugar box in Fig. 1, while still maintaining the flexibility to learn complicated multi-modal distribution cause by symmetry, as is the case with the bowl or ambiguity cause by similar viewpoints, as seen with the tuna can.

Although the feature comparison function $f(\phi_j|\phi)$ can be parameterized in a variety of ways, we parameterize it as a neural network that takes concatenated features ϕ and ϕ_j as input. Implementation details of our specific architecture and training procedure can be found in Section V-C.

V. Experimental Evaluation

A. Baselines

We compare our method to other common distribution estimation approaches.While the set of methods we compare to is far from exhaustive, we believe it represents a good sampling of the major classes of distribution estimation algorithms.

*1) **Fixed Isotropic Bingham**:* Given a base pose estimator (such as [8], [7]) which outputs a single point estimate $\bar{\mathbf{q}}$ of the object's orientation, a simple baseline method for estimating an orientation distribution is to fit a Bingham centered about $\bar{\mathbf{q}}$, with a fixed isotropic concentration parameter, σ. This parameter can be tuned independently for each object, using cross-validation. In our experiments, we fit this parameter using a sub-random search [41] over a validation set, maximizing the log likelihood of the ground truth orientation.

Note that, unlike our method described in Section IV, the uncertainty of this baseline does not depend on the input image; rather, a single uncertainty parameter is used for all images of a given object type. Thus, this approach is not sensitive to the uncertainties that can be induced by environmental factors such as lighting, viewpoint, or occlusions. We show that this approach performs significantly worse than our method which outputs image-dependent uncertainty estimates.

*2) **Mixture of Isotropic Binghams**:* Some methods, such as DenseFusion [8], output a set of orientation estimates $\mathbf{q}_i$, each with a corresponding confidence c_i. A mixture of isotropic Bingham distributions can be fit to this output, with each isotropic Bingham distribution centered at the orientation estimate $\mathbf{q}_i$ with a fixed concentration parameter, σ, similarly tuned using cross-validation. These Bingham distributions are combined into a single mixture distribution by weighting each one by its confidence c_i, where the confidence scores are normalized to sum to one.

*3) **MC-Dropout Ensemble**:* Monte Carlo Dropout [29] has been used to approximate the epistemic uncertainty of a network's predictions, using dropout to simulate an ensemble of estimators. PoseCNN [7] includes a dropout layer, whereas we retrained DenseFusion [8] with an additional dropout layer inserted into the network. At test time, n forward passes of the network are run on each observation, with dropout active, to generate n orientation estimates for each input. This process generates an estimate of the epistemic uncertainty and is mathematically equivalent to a deep Gaussian process [29]. We make the assumption that these samples are drawn from a Bingham distribution and fit the parameters of such a distribution to the sampled orientation estimates. The number of forward passes used provides a trade-off between the accuracy of the uncertainty estimates and the speed of computation; following previous work [42], we choose $n = 50$ as a balance between accuracy and speed.

*4) **Confusion Matrix**:* As described in [23], a confusion matrix can be used to estimate the conditional uncertainty $p(\mathbf{q}^* \mid \hat{\mathbf{q}})$ of an estimate $\hat{\mathbf{q}}$. The confusion matrix is computed over a discretization of the orientation space. This method counts how often the ground-truth orientation $\mathbf{q}^*$ is classified as $\hat{\mathbf{q}}$ by the our base estimator in a training or validation set. As with our method, we use the orientation discretization from Straub [40] to define the discretization of the confusion matrix.

Specifically, we form a $n \times n$ matrix, $\mathbf{X}$, where n is

the number of orientations in our discretization. Each row represents the estimated poses $\hat{q}_j$, whereas each column represents the ground-truth poses q^*. We initialize this matrix to 0. To compute the elements of this matrix, we iterate over our dataset. For each image I_j, we compute an estimated orientation $\hat{q}_j$ with a base pose estimator (e.g. [7] or [8]). Given the ground-truth pose q^*, we then increment the value of the matrix corresponding to the row and column of $(\hat{q}_j, q^*)$. A small constant ε is to each element of the confusion matrix for Laplace smoothing, and the rows are then normalized using the procedure described in Section IV-B.

At inference time, we compute the estimated orientation $\hat{\mathbf{q}}$ using the base estimator. The row in the confusion matrix that corresponds to this estimated orientation gives the estimated value of the distribution $p(\mathbf{q}^* \mid \hat{\mathbf{q}})$.

*5) **Full Bingham Regression**:* Using the parameterization described in Section IV-A, we can regress the parameters of a full Bingham distribution. We still require that the Bingham be centered at the output of the estimator, $\bar{\mathbf{q}}$, but the covariance can be dilated and rotated about this point. The four parameters of the diagonal concentration matrix, $\mathbf{Z}$, can be simplified to three parameters by subtracting the maximum value, without loss of generality [33]. To rotate the distribution about $\bar{\mathbf{q}}$, the 4D rotation matrix $\mathbf{M}$, can be post-multiplied by the four dimensional rotation matrix $\mathbf{Q}$, using a three dimensional rotation $\mathbf{R}_P$ parameterized by the quaternion $\mathbf{q}_P$, $\mathbf{Q} = diag\left(\begin{bmatrix}1 & \mathbf{R}_P\end{bmatrix}\right)$.

*6) **Direct Histogram Regression**:* As mentioned previously, we test directly regressing from the features $\phi(I)$ to the histogram values at each gridded orientation $\mathbf{q}_j$, as opposed to computing these values based on feature comparisons. For this baseline, the values at each grid cell, $p(\mathbf{q} \mid \phi)$, are estimated using a neural network, which receives as input the latent features ϕ and regresses an unnormalized posterior, $\hat{p}(\mathbf{q}_j \mid \phi)$. As before, we train this function with the log likelihood loss of equation 1. Also as before, we normalize over all of the gridded orientations, and use the gridding from Straub [40].

*7) **Cosine Feature Difference**:* As an ablation of our learned comparison method from Section IV-B, we evaluate using the cosine distance as the feature comparison function, e.g. $f(\phi_j, \phi) = \phi_j \cdot \phi / (||\phi_j|| ||\phi||)$. For this ablation, the cosine distance replaces our learned comparison function, to evaluate the benefits to learning such a comparison function. This distance function $f(\phi_j, \phi)$ is used to approximate $\hat{p}(\mathbf{q}_j \mid \phi)$ in the same manner as described in Section IV-B.

B. Dataset

To evaluate the accuracy of our methods for uncertainty estimation as well as the baselines, we use the YCB Video dataset [7], a commonly used pose estimation dataset. This dataset is comprised of videos of 21 objects in various cluttered tabletop scenes, with segmentation and 6D pose annotations. Each object in the dataset is accompanied by a textured mesh. Among the 21 objects, four objects contain discrete rotational symmetries, meaning the objects have a rotational symmetry with respect to a discrete set of rotations. One object (the bowl) has a continuous rotational symmetry, having a symmetric axis about which the object can be freely rotated. Twelve of the videos are held out as a test set, leaving 80 videos for training. We focus on this dataset for our evaluation, as the two base estimators that we evaluate, DenseFusion [8] and PoseCNN [7], have made the pretrained weight for these objects available.

C. Implementation Details

We tested each method for estimating orientation distributions using both PoseCNN [7] and DenseFusion [8] features. When generating features with DenseFusion, we used the segmentation estimated by PoseCNN for training images, as is done in the original publication [8] and the ground truth segmentation for the rendered images used for our non-parametric distributions. We use the global feature produced by DenseFusion for our multi-modal methods, while the maximum confidence local feature is used in our Bingham Regression method. These were experimentally verified to produce the best results in each method. All features are generated using pretrained models without further fine-tuning.

For PoseCNN features, we use the output of the last hidden layer of the network's orientation head. When generating PoseCNN features for rendered images, it is possible for the estimator to not detect the target object, as the network jointly estimates a segmentation mask as well as the pose of the object. In these cases, we evaluated each method using the featurization of the detected object whose mask maximally overlaps the target object. When the estimator failed to find any object in an image, we set the feature to the zero vector. This process is only used for rendered images. For real images, only the features of objects detected by PoseCNN are used.

Our methods are trained using a combination of real and rendered data. This data is resampled to ensure a uniform coverage over SO(3) using the discretization method described in Section III-B. In this case, we use coarser discretization than our distribution gridding, with a maximum distance to the nearest bin center of about 26 degrees.

Our non-parametric methods used a simple three layer neural network with 4096 neurons on each hidden layer, dropout and ReLU activations on the input and first hidden layer, and sigmoid activation on the output. The parametric methods draw inspiration from DenseFusion [8], using four fully connected layers, with 640, 256, and 128 neurons on the hidden layers and ReLU activation functions.

D. Evaluation Method

We evaluate each orientation distribution estimator on each example in the YCB test set and record the log likelihood of the ground-truth pose, clipped to a minimum of 1e-6. A likelihood distribution is computed for each of these images and the likelihood of the ground truth pose is computed given that distribution. For multi-modal methods, the interpolation described in Section IV-B is used, while Bingham based methods use standard Bingham likelihood. The log

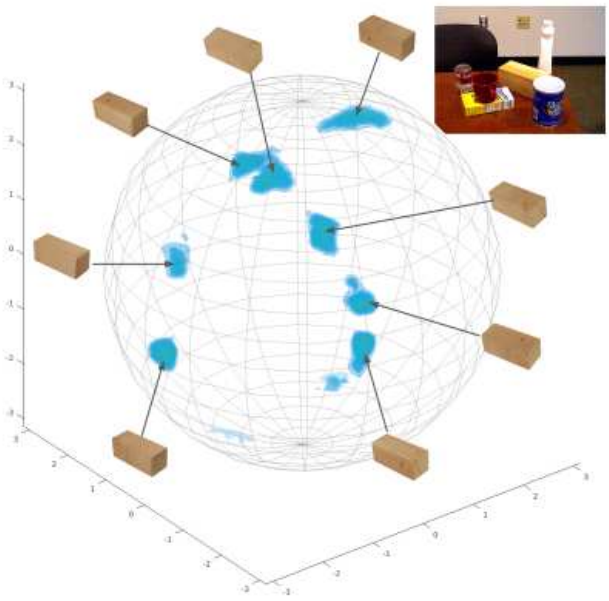

Fig. 4. Multimodal distribution of the wood block's symmetries captured by the Learned Comparison Estimator, using PoseCNN features. There are eight distinct modes, associated with four 90 degree rotations about the long axis multiplied by two 180 degree rotations about one of the short axes. This distribution is impossible to well model with a single Bingham distribution, as shown in Fig. 3, but can be easily captured by a multi-modal histogram.

likelihood evaluation metric allows us to evaluate whether the distribution is correctly placing probability mass in the appropriate locations.

VI. Results

The log likelihood results of our method and all the baselines can be seen in Table I. We separate the objects into symmetric and non-symmetric objects and evaluate each method using DenseFusion [8] and PoseCNN [7] features. We find that our method of isotropic Bingham regression performs the best for non-symmetric objects when combined with DenseFusion features. Good performance is also obtained with a Bingham distribution fit to samples from MC Dropout using PoseCNN features. A uni-modal Bingham distribution is able of capture the orientation uncertainty of non-symmetric objects when the distribution is tightly fit around a mean orientation, as shown by the tomato soup can in Fig. 3. However, such a method will struggle with symmetric objects, like the wooden block in Fig. 3, or objects that appear symmetric from particular views or under particular occlusion patterns.

While the Full Bingham Regression performed similarly to the Isotropic Bingham Regression, we found this method to be less numerically stable in training, as it requires the gradients for the normalization constant of an anisotropic Bingham distribution. The gradients of the isotropic normalization constant, however, proved to be more stable and cause few problems in training. Our experiments demonstrate that this longer training time provides little benefit over the isotropic version.

For symmetric objects, Table I shows that learning a non-parametric histogram distribution is best able to capture the multi-modal nature of the uncertainty of such objects. Specifically, Table I shows that our Learned Comparison Histogram estimation method has the best log likelihood, when using PoseCNN features. PoseCNN features using Histogram Regression is also among the top scoring methods for this task, although performance is significantly worse than our method. Note that the log likelihoods of the symmetric objects are expected to be lower than the log likelihood for non-symmetric objects, since the optimal distribution will spread the probability mass evenly over each symmetric mode, leading to a lower likelihood at each mode. This can be seen when our method correctly distributes the probability density to all eight of the wood block's symmetric modes, shown in Fig. 4. Overall, our learned comparison based method for estimating a non-parametric distribution is best able to capture the uncertainty across the full set of objects, having the flexibility to model multi-modal distributions for objects with various types of symmetries, while still being able to concentrate the probability mass over a single mode when necessary.

We note that the log likelihood values in Table I may be hard for the reader to interpret directly; for reference, a uniform distribution, where every orientation is equally likely, would be expected to obtain a log likelihood of -2.29. As shown in Table I, some distributions perform worse than the uniform distribution. This is likely caused by overestimating the certainty of the output, i.e. the distribution for such methods is often concentrated around a single incorrect mode. In such cases, the method fails to put sufficient probability mass in regions of the pose space far from this incorrect mode, leading to a very low log likelihood at the ground-truth pose.

Table I also reveals that DenseFusion performs poorly on uncertainty estimation for symmetric objects, for all methods and baselines. Our analysis revealed that this is due to DenseFusion's lack of robustness to poor segmentation masks. To demonstrate this, we evaluated our Learned Comparison method using DenseFusion features but using ground truth masks, instead of estimated masks. The results, shown in Table II, reveal a substantial increase in performance for the log likelihood of symmetric objects, when using ground truth masks instead of estimated masks. This experiment reveals the large contribution of poor segmentation to the overall pose uncertainty in Table I, for DenseFusion on symmetric objects. In contrast, because PoseCNN does not receive as input a segmentation mask, it is more robust to these types of errors.

A. Confidence Filtering

As previously shown [28], pose uncertainty estimation can be used to robustly filter pose estimates. As we are directly computing the likelihood of an estimate, the output of our algorithm can be used to select which poses to trust and which to reject. Specifically, we use each of our methods to estimate a distribution over orientations. We then compute a

	Our Method		Baselines						
Objects	Bingham Regression	Learned Comparison	Fixed Bingham	Bingham Mixture	Dropout	Confusion Matrix	Cosine Distance	Full Bingham	Histogram Regression
Non-Symmetric									
DenseFusion	**2.80**	1.18	1.74	0.66	0.70	1.63	-1.90	2.56	0.28
PoseCNN	1.91	2.17	1.50	-	*2.71*	-2.46	-0.92	1.95	1.87
Symmetric									
DenseFusion	-3.81	-5.54	-3.66	-2.27	-8.09	-2.91	-2.23	-4.18	-2.57
PoseCNN	-8.82	**-0.52**	-9.18	-	-5.28	-7.75	-1.55	-3.70	*-1.23*
All									
DenseFusion	*1.72*	0.08	0.86	0.18	-0.74	0.88	-1.95	1.46	-0.19
PoseCNN	0.19	**1.74**	-0.22	-	1.43	-3.31	-1.02	1.05	1.37

TABLE I

MEAN LOG LIKELIHOOD OF GROUND TRUTH ORIENTATION. FOR EACH GROUPING, BEST-SCORING METHODS ARE MARKED IN BOLD; SECOND-BEST SCORING METHODS ARE INDICATED BY ITALICS.

	Non-Symmetric	Symmetric	All
Estimated Masks	1.18	-5.54	-0.18
Ground Truth Masks	1.97	-0.18	1.61

TABLE II

MEAN LOG LIKELIHOOD OF GROUND TRUTH ORIENTATION FOR LEARNED COMPARISON ESTIMATOR USING DENSEFUSION FEATURES WITH ESTIMATED AND GROUND TRUTH MASKS.

pose estimate $\hat{\mathbf{q}}$ from the base pose estimator, and we use our estimated distributions to compute the likelihood at this pose: $p(\hat{\mathbf{q}} \mid \phi(I))$. For our Learned Comparison method, this requires interpolating the histogram, which we achieve using the interpolation described in Section IV-B.

We test the validity of this process in Table III, which shows the effects of rejecting pose estimates based on likelihood thresholds. In this experiment, we describe these thresholds as multiples of the likelihood of a sample selected at from a uniform distribution, 0.101. As a reminder, this is a probability density, rather than a discrete probability value, and thus ranges from 0 to infinity. For the remaining poses, angular error is calculated with respect to annotated symmetry axes and Average Distance Error (ADD) and Symmetric Average Distance Error (ADD-S) is computed for non-symmetric objects and symmetric objects, respectively. Further details on these evaluation metrics can be found in prior works [7], [8], [28].

Our results can be seen in Table III, which shows a clear trend of decreasing angular error with an increasing threshold of estimated log likelihood. This shows that using a threshold on the estimated log likelihood (using our methods for estimating orientation distributions) is indeed an effective approach for filtering out examples with a large angular error. Such a threshold can be used to allow a robot to determine when its predictions might be inaccurate. In such cases, the robot can move its camera to acquire new viewpoints before taking an action, or it can ask a human for help.

Learned Comparison (PoseCNN)

Threshold	Ang Error (deg)	ADD (m)	Reject (%)
-	25.44	0.0402	0
Uniform	24.76	0.0398	3
10x Uniform	23.69	0.0390	7
50x Uniform	17.12	0.0374	20
100x Uniform	15.90	0.0361	34
200x Uniform	12.72	0.0364	71

Bingham Regression (DenseFusion)

Threshold	Ang Error (deg)	ADD (m)	Reject (%)
-	21.68	0.0155	0
Uniform	21.61	0.0155	0
50x Uniform	19.08	0.0145	11
250x Uniform	16.91	0.0135	18
1e3x Uniform	13.74	0.0118	25
2e3x Uniform	12.53	0.0112	30

(a) Non-Symmetric Objects

Learned Comparison (PoseCNN)

Threshold	Ang Error (deg)	ADD-S (m)	Reject (%)
-	40.05	0.0478	0
Uniform	34.13	0.0472	13
2x Uniform	32.60	0.0475	16
5x Uniform	29.24	0.0468	24
15x Uniform	25.43	0.0487	40

(b) Symmetric Objects

TABLE III

POSE ERROR COMPUTED ON ESTIMATES BELOW LIKELIHOOD THRESHOLDS FOR NON-SYMMETRIC (A) AND SYMMETRIC (B) OBJECTS. THE THRESHOLDS ARE DESCRIBED AS MULTIPLES OF CHANCE, THE LIKELIHOOD OF A UNIFORM DISTRIBUTION (0.101).

VII. CONCLUSION

We propose two methods for augmenting existing pose estimation methods with orientation distributions. These methods were compared to a series of uncertainty estimation baselines, evaluated using the log likelihood of the ground-truth orientation. Our findings indicate that, for non-symmetric objects, our learned isotropic Bingham regression gives the best performance. For objects with unknown symmetries, our method for estimating a non-parametric distribution based on a learned feature comparison gives the best performance. We demonstrate that our method can be used to filter out the examples with the worst angular error, for which the robot can choose to capture more information about the environment or request help from a human. Future work will use this uncertainty estimation in the context of tracking or grasping applications; we will also explore how multiple methods for estimating uncertainty can be combined for improved performance.

REFERENCES

[1] S.-K. Kim and M. Likhachev, "Planning for grasp selection of partially occluded objects," in *ICRA*, 2016.

[2] N. T. Dantam, Z. K. Kingston, S. Chaudhuri, and L. E. Kavraki, "Incremental task and motion planning: A constraint-based approach." in *RSS*, 2016.

[3] G. Thomas, M. Chien, A. Tamar, J. A. Ojea, and P. Abbeel, "Learning robotic assembly from cad," in *ICRA*, 2018.

[4] C. Goldfeder, M. Ciocarlie, H. Dang, and P. K. Allen, "The columbia grasp database," in *ICRA*, 2009.

[5] M. Ciocarlie, K. Hsiao, E. G. Jones, S. Chitta, R. B. Rusu, and I. A. Şucan, "Towards reliable grasping and manipulation in household environments," in *Experimental Robotics*, 2014.

[6] M. Zucker, N. Ratliff, A. D. Dragan, M. Pivtoraiko, M. Klingensmith, C. M. Dellin, J. A. Bagnell, and S. S. Srinivasa, "Chomp: Covariant hamiltonian optimization for motion planning," *IJRR*, 2013.

[7] Y. Xiang, T. Schmidt, V. Narayanan, and D. Fox, "Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes," 2018.

[8] C. Wang, D. Xu, Y. Zhu, R. Martín-Martín, C. Lu, L. Fei-Fei, and S. Savarese, "Densefusion: 6d object pose estimation by iterative dense fusion," in *CVPR*, 2019.

[9] E. Brachmann, A. Krull, F. Michel, S. Gumhold, J. Shotton, and C. Rother, "Learning 6d object pose estimation using 3d object coordinates," in *ECCV*, 2014.

[10] W. Kehl, F. Manhardt, F. Tombari, S. Ilic, and N. Navab, "Ssd-6d: Making rgb-based 3d detection and 6d pose estimation great again," in *ICCV*, 2017.

[11] J. Tremblay, T. To, B. Sundaralingam, Y. Xiang, D. Fox, and S. Birchfield, "Deep object pose estimation for semantic robotic grasping of household objects," *arXiv preprint arXiv:1809.10790*, 2018.

[12] J. M. Wong, V. Kee, T. Le, S. Wagner, G.-L. Mariottini, A. Schneider, L. Hamilton, R. Chipalkatty, M. Hebert, D. Johnson, J. Wu, B. Zhou, and A. Torralba, "Segicp: Integrated deep semantic segmentation and pose estimation," 2017.

[13] A. Zeng, K.-T. Yu, S. Song, D. Suo, E. Walker, A. Rodriguez, and J. Xiao, "Multi-view self-supervised deep learning for 6d pose estimation in the amazon picking challenge," in *ICRA*, 2017.

[14] M. A. Fischler and R. C. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *Communications of the ACM*, 1981.

[15] S. Hinterstoisser, V. Lepetit, S. Ilic, S. Holzer, G. Bradski, K. Konolige, and N. Navab, "Model based training, detection and pose estimation of texture-less 3d objects in heavily cluttered scenes," in *ACCV*, 2012.

[16] Z. Cao, Y. Sheikh, and N. K. Banerjee, "Real-time scalable 6dof pose estimation for textureless objects," in *ICRA*, 2016.

[17] T. Hodaň, X. Zabulis, M. Lourakis, Š. Obdržálek, and J. Matas, "Detection and fine 3d pose estimation of texture-less objects in rgb-d images," in *IROS*, 2015.

[18] B. Drost, M. Ulrich, N. Navab, and S. Ilic, "Model globally, match locally: Efficient and robust 3d object recognition," in *CVPR*, 2010.

[19] M. Sundermeyer, Z.-C. Marton, M. Durner, M. Brucker, and R. Triebel, "Implicit 3d orientation learning for 6d object detection from rgb images," in *ECCV*, September 2018.

[20] P. Wohlhart and V. Lepetit, "Learning descriptors for object recognition and 3d pose estimation," in *CVPR*, 2015.

[21] V. Balntas, A. Doumanoglou, C. Sahin, J. Sock, R. Kouskouridas, and T.-K. Kim, "Pose guided rgbd feature learning for 3d object pose estimation," in *ICCV*, 2017.

[22] H. Su, C. R. Qi, Y. Li, and L. J. Guibas, "Render for cnn: Viewpoint estimation in images using cnns trained with rendered 3d model views," in *ICCV*, 2015.

[23] Z. C. Márton, S. Türker, C. Rink, M. Brucker, S. Kriegel, T. Bodenmüller, and S. Riedel, "Improving object orientation estimates by considering multiple viewpoints," *Autonomous Robots*, 2018.

[24] J. Glover, G. Bradski, and R. B. Rusu, "Monte carlo pose estimation with quaternion kernels and the bingham distribution," in *RSS*, 2012.

[25] S. Riedel, Z.-C. Marton, and S. Kriegel, "Multi-view orientation estimation using bingham mixture models," in *AQTR*, 2016.

[26] E. Brachmann, F. Michel, A. Krull, M. Ying Yang, S. Gumhold, and C. Rother, "Uncertainty-driven 6d pose estimation of objects and scenes from a single rgb image," in *CVPR*, 2016.

[27] X. Deng, A. Mousavian, Y. Xiang, F. Xia, T. Bretl, and D. Fox, "Poserbpf: A rao-blackwellized particle filter for 6d object pose estimation," in *RSS*, 2019.

[28] F. Manhardt, D. M. Arroyo, C. Rupprecht, B. Busam, T. Birdal, N. Navab, and F. Tombari, "Explaining the ambiguity of object detection and 6d pose from visual data," in *ICCV*, 2019.

[29] Y. Gal and Z. Ghahramani, "Dropout as a bayesian approximation: Representing model uncertainty in deep learning," in *ICML*, 2016.

[30] A. Kendall and Y. Gal, "What uncertainties do we need in bayesian deep learning for computer vision?" in *Advances in Neural Information Processing Systems*, 2017.

[31] R. A. Srivatsan, M. Xu, N. Zevallos, and H. Choset, "Bingham distribution-based linear filter for online pose estimation," in *RSS*, 2017.

[32] I. Gilitschenski, G. Kurz, S. J. Julier, and U. D. Hanebeck, "Unscented orientation estimation based on the bingham distribution," *IEEE Transactions on Automatic Control*, 2015.

[33] C. Bingham, "An antipodally symmetric distribution on the sphere," *The Annals of Statistics*, 1974.

[34] B. Grossmann and V. Krüger, "Fast view-based pose estimation of industrial objects in point clouds using a particle filter with an icp-based motion model," in *INDIN*, 2017.

[35] K. Shoemake, "Uniform random rotations," in *Graphics Gems III (IBM Version)*, 1992.

[36] A. Yershova, S. Jain, S. M. Lavalle, and J. C. Mitchell, "Generating uniform incremental grids on so (3) using the hopf fibration," *The International journal of robotics research*, vol. 29, no. 7, 2010.

[37] X. Perez-Sala, L. Igual, S. Escalera, and C. Angulo, "Uniform sampling of rotations for discrete and continuous learning of 2d shape models," in *Robotic vision: Technologies for machine learning and vision applications*, 2013, pp. 23–42.

[38] E. B. Dam, M. Koch, and M. Lillholm, *Quaternions, interpolation and animation*, 1998, vol. 2.

[39] A. Perez-Gracia and F. Thomas, "On cayleys factorization of 4d rotations and applications," *Advances in Applied Clifford Algebras*, 2017.

[40] J. Straub, T. Campbell, J. P. How, and J. W. Fisher III, "Efficient global point cloud alignment using bayesian nonparametric mixtures," in *CVPR*, 2017.

[41] O. Bousquet, S. Gelly, K. Kurach, O. Teytaud, and D. Vincent, "Critical hyper-parameters: No random, no cry," *arXiv preprint arXiv:1706.03200*, 2017.

[42] A. Kendall, V. Badrinarayanan, and R. Cipolla, "Bayesian segnet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding," *arXiv preprint arXiv:1511.02680*, 2015.