Ki-Pode: Keypoint-based Implicit Pose Distribution Estimation of Rigid Objects

Thorbjørn Mosekjær Iversen
thmi@mmmi.sdu.dk
Rasmus Laurvig Haugaard
rlha@mmmi.sdu.dk
Anders Glent Buch
anbu@mmmi.sdu.dk

SDU Robotics
The Maersk Mc-Kinney Moller Institute
University of Southern Denmark, DK

Abstract

The estimation of 6D poses of rigid objects is a fundamental problem in computer vision. Traditionally pose estimation is concerned with the determination of a single best estimate. However, a single estimate is unable to express visual ambiguity, which in many cases is unavoidable due to object symmetries or occlusion of identifying features. Inability to account for ambiguities in pose can lead to failure in subsequent methods, which is unacceptable when the cost of failure is high. Estimates of full pose distributions are, contrary to single estimates, well suited for expressing uncertainty on pose. Motivated by this, we propose a novel pose distribution estimation method. An implicit formulation of the probability distribution over object pose is derived from an intermediary representation of an object as a set of keypoints. This ensures that the pose distribution estimates have a high level of interpretability. Furthermore, our method is based on conservative approximations, which leads to reliable estimates. The method has been evaluated on the task of rotation distribution estimation on the YCB-V and T-LESS datasets and performs reliably on all objects.

1 Introduction

Many robotics applications which involve the manipulation of rigid objects require accurate knowledge of object poses to succeed. The topic of pose estimation has, thus, received considerable attention from the computer vision community. While the literature on pose estimation covers many different sensor modalities, pose estimation from an RGB image remains highly relevant due to the low cost of industry-grade RGB cameras.

The majority of the literature on pose estimation is concerned with the estimation of a single best estimate of object pose. However, a single estimate is inherently ill-suited to express visual ambiguity which is unavoidable in many practical applications due to occlusions, poor image and lighting conditions, or object symmetries. In recent years there has, therefore, been an increasing interest in the estimation of probability distributions over object pose.

Distribution estimates are especially valuable when used in contexts with a high cost of failure such as robotic grasping. In such contexts, it is favorable to postpone a task until the task's probability of success is above a certain threshold. While a rough estimate of task success can be computed from a single pose estimate with associated variance, a full pose distribution allows for more advanced probability-based inference of task success. Overconfident estimates of uncertainty can, however, still lead to failure, so it is crucial that the estimates are reliable, in the sense that the false positive rate is low. Underconfident estimates simply risk being so uncertain that subsequent tasks are not performed, even when the information necessary for successful completion was available. However, this can be alleviated through sensor fusion with additional views which the distribution estimates are well suited for.

Estimating a full 6D probability distribution over object pose is non-trivial, and most prior work focuses on rotation distribution estimation, modeling the uncertainty either as a mixture of parametric distributions or as a probability mass function over sampled rotations in SO(3). We argue that both methods are somewhat limited in their descriptive ability, restricted by the number of distributions or the number of samples, respectively.

Our contribution is a pose distribution estimation method that estimates a probability density function over object pose from a single RGB image. The function is formulated implicitly which enables the predicted distribution to be continuous without being constrained by the descriptive ability of parametric distributions. The formulation is based on an explicit derivation of the probability density function using object keypoints as an intermediary representation of object pose. This increases the interpretability of the estimates as illustrated in Fig. 1, which shows estimated keypoint distributions and resulting pose distribution for three different objects. Our method is based on conservative approximations which ensures a high level of reliability.

The scope of the evaluation is limited to rotation distribution estimation. This is done to allow for comparison to existing methods, which primarily focus on estimating the distribution on SO(3). Our method is evaluated on the YCB-V and T-LESS datasets and compared to previously published results of state-of-the-art methods.

The paper is organized as follows: Section 2 presents related work, followed by section 3 in which our method is presented. The implementation of the method together with the experimental procedure is presented in section 4, and results are shown and discussed in section 5. Finally section 6 presents the conclusion.

2 Related work

The estimation of object pose from a single RGB image is a topic that has received extensive study by the machine vision community. Our method is inspired by single pose estimation methods that rely on locating object keypoints in images through the estimation of heatmaps, which encode the probability distribution of the keypoints, e.g. [13].

When object keypoints are easily recognizable, such that their uncertainty distribution in the image plane is approximately Gaussian, it is possible to analytically propagate the uncertainty from keypoints to pose [5]. However, the assumption of normality becomes very poor when visual ambiguities are present.

One approach to estimating a pose distribution is by representing the distribution as an ensemble of pose estimates from one or more pose estimation methods [9, 14, 17]. These methods rely on the assumption that the ensemble accurately represents the true pose distribution. However, unless the pose estimators generating the ensemble are explicitly trained for this, it is questionable if the estimated pose distribution reflects the true distribution.

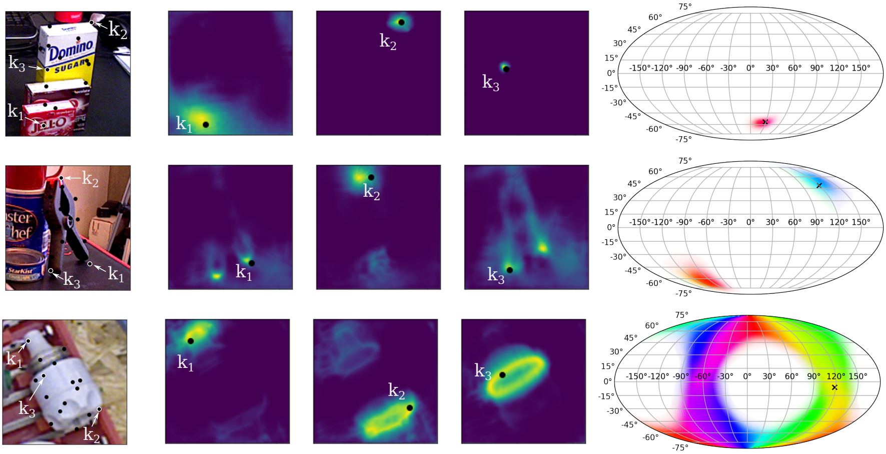

Figure 1: Visualization of an image crop, the logarithm of the marginal probability distributions for the first three keypoints, and a visualization of the resulting orientation distribution. The rows shows the YCB-V sugar box, YCB-V large clamp, and T-LESS object 1.

Pose distributions are often multimodal and difficult to model analytically. Therefore, it is common to assume that the rotational and positional part of the distribution is decoupled, assume that the object position is normally distributed, and focus on object rotation estimation. A common approach to rotation distribution estimation is to sample SO(3) and attribute a probability to each unique rotation. This has been done using a confusion matrix constructed from training data [10], treating rotation estimation as a discrete classification problem [18], or training a network to compute latent features which implicitly encode orientation such that comparison between sampled rotations and a query image is done in latent space. This has both been done with an autoencoder using cosine distance to measure latent feature similarity [3, 19], or using the latent features from an existing pose estimator combined with an MLP network for regressing unnormalized likelihood [14].

An alternative to the sampling approach is to model the pose distribution using parametric distributions. Although it has been argued that the Matrix Fisher distribution is well suited for deep learning-based parameter regression [11], the most common choice of orientation distribution is the Bingham distribution since it explicitly handles the antipodal symmetry of unit quaternions, which are often used to parameterize orientation. Since unimodal distributions are unable to express multiple orientation ambiguities, several works estimate the pose uncertainty as a mixture of Binghams. This has been done by fitting Bingham distributions to the output of a multiple hypotheses pose estimator [9], or direct regression of the Bingham parameters using deep learning [2, 6, 14]. Related is [15] in which a network is trained to estimate 1D rotation distributions as mixtures of von Mises distributions.

Similar to our method, [12] proposes to express the probability density function implicitly. In their work, a network is trained end-to-end to output the unnormalized likelihood from an image and a query pose. The likelihood function is then normalized using an equivolumetric grid. Unlike [12], our method uses an intermediary representation which increases the interpretability of the pose distribution estimates. Furthermore, our intermediary representation is translation equivariant, making it automatically generalize to arbitrary image translations, which is arguably the most important regularization on the image domain.

3 Method

Our method uses object keypoints as a means to estimate object pose. N keypoints are defined in the object's reference frame and a network is trained to estimate the marginal probability distributions over the keypoints' projections on the image plane. These marginal distributions will be referred to as heatmaps. The pose distribution is formulated implicitly, in that given a query pose and the heatmaps, the method returns the unnormalized likelihood of the given pose. The normalized likelihood can then be computed using a sampling-based strategy.

The following three subsections discuss the details of our method. Section 3.1 derives and discusses the relationship between pose distribution and image keypoints as well as the approximations needed when using marginal instead of conditional probabilities. Section 3.2 discusses the training of the network which estimates the heatmaps. Finally, section 3.3 discusses the sampling-based method of normalization.

3.1 Relationship between pose distribution and keypoints

Pose estimation from a single RGB camera is defined as the estimation of an object's pose, θ, which consists of the rotation, $R \in \mathrm{SO}(3)$, and translation, $t \in \mathbb{R}^3$, relative to the camera from a single RGB image I. Pose distribution estimation can then be defined as the estimation of a probability density function over object pose conditioned on the input image, $\mathrm{p}(\theta|I)$.

In the proposed method N object keypoints are used to model the correspondence between an object and its projection on the image plane. The keypoints, $K_i \in \mathbb{R}^3$, are defined relative to the object reference frame. Given an object pose and camera intrinsics the 3D keypoints can be projected onto the image plane resulting in N 2D keypoints, $k_i \in \mathbb{R}^2$. Using the law of total probability, the keypoints as an intermediary representation, and the chain rule, the pose distribution can be reformulated as:

$$\mathrm{p}(\theta|I) = \int \mathrm{p}(\theta, \mathbf{k}|I) d\mathbf{k} = \int \mathrm{p}(\theta|\mathbf{k}, I)\mathrm{p}(\mathbf{k}|I) d\mathbf{k} \tag{1}$$

where $p(\mathbf{k}|I)$ is the joint distribution over k_i for $i = 1, 2..., N$ conditioned on the image, and $p(\theta|\mathbf{k}, I)$ is the distribution over object pose conditioned on both the image and a fixed keypoint configuration. A set of four or more non-colinear keypoints uniquely define an object pose, so the likelihood $p(\theta|\mathbf{k}, I)$ will only be non-zero for the unique pose where the projection of keypoints $\mathbf{K}$ on the image plane is identical to the keypoint configuration $\mathbf{k}$. This can be expressed using the Dirac delta function as:

$$\mathrm{p}(\theta|\mathbf{k}, I) = \delta(\mathbf{k}_\theta - \mathbf{k}) \tag{2}$$

where $\mathbf{k}_\theta$ is the projection of the keypoints $\mathbf{K}$ using the camera intrinsics and the pose θ. Using Eq. 2 the integral in Eq. 1 can be rewritten as:

$$\mathrm{p}(\theta|I) = \int \delta(\mathbf{k}_\theta - \mathbf{k})\mathrm{p}(\mathbf{k}|I) d\mathbf{k} = \mathrm{p}(\mathbf{k}_\theta|I) \tag{3}$$

Eq. 3 states that the likelihood of the object being in pose θ is equal to the joint likelihood of observing the keypoint configuration $\mathbf{k}_\theta$.

While estimating the joint distribution over keypoints $\mathrm{p}(\mathbf{k}|I)$ is non-trivial, the estimation of the marginal likelihoods, $\mathrm{p}(k_\mathrm{i}|I)$ are more easily achieved. A common approximation of the joint distribution from marginal distributions is the assumption of independence:

$$\mathrm{p}(\mathbf{k}) \approx \mathrm{p}(k_1)\mathrm{p}(k_2)...\mathrm{p}(k_N) \tag{4}$$

However, when the assumption of independence is invalid the approximation can be very poor. The effect is most extreme when all keypoints are strongly coupled such that the value of a single keypoint uniquely defines the rest. If $\mathbf{k}_{\text{-i}}$ is used to denote all keypoints excluding k_i, the chain rule states that $\mathrm{p}(\mathbf{k}) = \mathrm{p}(\mathbf{k}_{\text{-i}}|k_i)\mathrm{p}(k_i)$. Under strong coupling and assuming a valid configuration $\mathbf{k}$, this simplifies to $\mathrm{p}(\mathbf{k}) = \mathrm{p}(k_i)$. The product of the marginals in the extreme case thus yields $\mathrm{p}(k_1)\mathrm{p}(k_2)...\mathrm{p}(k_N) = \mathrm{p}(\mathbf{k})^N$. Consequently, under strong coupling, the joint distribution can be computed from the marginals as:

$$\mathrm{p}(\mathbf{k}) = \sqrt[N]{\mathrm{p}(k_1)\mathrm{p}(k_2)...\mathrm{p}(k_N)} \tag{5}$$

Eq. 4 and 5 express two extremes. The error of using 4 when the keypoints are strongly coupled is to diminish regions of low probability while using 5 when the keypoints are loosely coupled will lead to an attenuation of regions of low probability. As discussed in Sec. 1, it is crucial that estimates are not overconfident, while underconfident estimates are acceptable. For this reason, our method uses the conservative approximation formulated in Eq. 5. Combining Eq. 3 and 5 yields:

$$\mathrm{p}(\theta|I) \approx \hat{\mathrm{p}}(\theta|I) = C\sqrt[N]{\prod_{i=1}^{N}\mathrm{p}(k_{\theta,i}|I)} \tag{6}$$

where C is a normalization constant.

A note on constraint relaxation

The one-to-one correspondence between a pose θ and the keypoint configuration $\mathbf{k}_\theta$ expressed in Eq. 2, may seem strict considering potential discrepancies between the real-world objects and their synthetic counterparts, on which we train. Theoretically, if the individual keypoints are very well localized in the heatmaps, but there are significant systematic errors due to the sim-to-real domain gap, it would be possible to encounter situations where there does not exist a pose θ where $\hat{\mathrm{p}}(\theta|I) \neq 0$. However, the Dirac delta constraint is implicitly relaxed when the heatmap distributions have a dispersion that is relatively large compared to the systematic errors from the discrepancies, so additional relaxation was not found to be required.

3.2 Keypoint heatmaps

The proposed method assumes that an object detector is available to provide an $r \times r$ crop of the object, where r is the crop resolution. It is assumed that this crop contains the entire object, both visible and occluded parts, such that all keypoints are inside the crop. The heatmaps are estimated from the crops using a U-Net [16], with an RGB image as input

$(r \times r \times 3)$ and the heatmaps as output $(r \times r \times N)$, where the i'th channel in the output is an estimate of $\mathrm{p}(k_i|I)$.

To learn the heatmaps $\mathrm{p}(k_i|I)$, we represent the problem as spatial classification using cross-entropy loss, where the i'th channel represents a discrete distribution over the $r \times r$ pixels, and the ground truth class is the nearest pixel to the projected keypoint k_i. Identification of the exact nearest pixel may encourage overfitting, so as to avoid this, we regularize the training by instead setting the ground truth heatmap to an isotropic Gaussian centered around the ground truth keypoint, $\mathrm{p}(k_i|I) = \mathcal{N}(k_i, \sigma)$. The total loss is then a sum over the N cross-entropy losses:

$$L = -\sum_{i=1}^{N} \mathcal{N}(k_i, \sigma) \log \hat{\mathrm{p}}(k_i|I) \tag{7}$$

3.3 Normalization

The normalization constant C in Eq. 6 can be approximated by densely sampling SE(3) and choosing C such that $\sum_{\theta_i \in SE(3)} \hat{\mathrm{P}}(\theta_\mathrm{i}|I) = 1$, where $\hat{\mathrm{P}}(\theta_\mathrm{i}|I)$ is used to denote the probability mass of sample i. However, even when the spatial dimensions are bounded, densely sampling SE(3) requires a very large number of samples, since with n samples along a dimension the computational complexity is $\mathrm{O}(n^6)$. For dense sampling to be feasible, the computation of likelihood must be both fast and parallelizable. Our method has been explicitly developed for this and is capable of computing the likelihood for millions of samples pr. second. This is possible because the image only has to be passed through the network once, after which the computation of likelihood for a single sample only requires the projection of the keypoints followed by a lookup in the heatmaps, all of which are done in parallel on a GPU.

The sampling of SO(3) is done using the method presented in [21] as suggested by [12]. The method decouples a rotation into the direction of the frame's z-axis expressed as a point on a 2D unit sphere, and a tilt around the z-axis. The unit sphere is sampled in an equiarea grid using the HEALPix method[7], while the sampling of the tilt is chosen to make the sides of each volume element equal. This creates an equivolumetric grid with each volume being $V_\mathrm{grid} = \pi^2/M$, where M is the number of samples in the grid. The sampling method divides SO(3) recursively and for s recursions the number of samples is $M = 72 \cdot 8^s$. Since the grid is equivolumetric, the likelihoods can be approximated from the probability masses using $\hat{\mathrm{p}}(\theta_\mathrm{i}|I) \approx \hat{\mathrm{P}}(\theta_\mathrm{i}|I)/V_\mathrm{grid}$.

4 Experiments

The implementation of the proposed method uses the U-Net implementation from [20], which uses a ResNet-18 backbone pretrained on ImageNet. The number of keypoints is chosen to be $N = 16$. The object keypoints $\mathbf{K}$ are chosen using furthest distance sampling, which ensures that the keypoints are spread out over the surface of the object. The model is trained using physically-based BlenderProc [4] renderings provided in the YCB-V and T-LESS parts of the BOP-challenge dataset [8]. The standard deviation of the Gaussian used in the loss function was set to 1px and the training was done for 10 epochs on a GeForce GTX 1080 Ti GPU. To avoid training on images where most of the object is outside the image or the object is occluded, the training is only performed on images for which the object appears

	Summary of Evaluation in [14]				
	I	II	III	IV	Ki-Pode (ours)
master chef can	-0.78	-1.66	*2.32*	0.09	**4.36**
cracker box	*4.03*	3.75	-0.09	3.71	**6.05**
sugar box	3.77	*5.94*	2.62	4.20	**6.47**
tomato soup can	0.90	2.02	2.52	*3.99*	**5.16**
mustard bottle	3.85	4.61	3.02	*4.81*	**5.45**
tuna fish can	-3.12	-0.20	*2.64*	1.23	**5.10**
pudding box	2.18	2.64	3.13	*4.54*	**4.91**
gelatin box	4.65	*6.25*	3.53	5.73	**6.35**
potted meat can	1.18	***3.28***	1.60	3.06	3.11
banana	2.58	0.58	2.27	*3.70*	**4.80**
pitcher base	3.34	4.68	2.35	*4.88*	**5.39**
bleach cleanser	3.91	***4.70***	2.29	3.38	4.02
bowl	***1.37***	-2.77	-1.21	-9.62	-1.50
mug	3.73	2.50	2.75	***4.72***	4.62
power drill	4.31	*5.92*	2.43	4.17	**6.14**
wood block	2.64	-2.09	-0.51	***4.49***	-1.76
scissors	*3.74*	0.51	0.66	1.63	**5.49**
large marker	-7.59	-0.29	*1.02*	-8.13	**3.94**
large clamp	-5.64	-6.67	-1.63	-3.54	**3.07**
extra large clamp	-5.20	-2.92	*0.19*	-5.03	**2.25**
foam brick	-0.26	-2.28	*1.70*	-12.0	**2.44**
All	0.86	1.71	1.74	1.43	**4.09**

Table 1: The table shows the mean log-likelihood score for our method as well as the top 4 performing pose distribution methods from [14]. The roman numerals refer to the methods I: Conf w/DenseFusion, II: Reg ISO w/DenseFusion, III: Comp w/PoseCNN, and IV: Dropout w/PoseCNN. For a description of each method see the original paper. Italic font is used to highlight the best performance among the evaluations in [14], while bold font highlights the best performance among all methods including ours. Red highlights scores below -2.29 which is worse than a uniform distribution.

in the image and is at maximum 95% occluded. During training, the images are augmented using random gamma, Gaussian blur, Gaussian noise, ISO noise, and color jitter [1], as well as random scaling, rotation, and translation of the image crop. Our implementation of the SO(3) grid sampling method is based on the one provided by [12].

The evaluation focuses on estimating the rotation distribution, $\hat{p}(R|I)$, such that the proposed method can be compared with existing works. Our method estimates an implicit distribution on SE(3), so the rotational part of the distribution must be marginalized by integrating over the spatial dimensions. The normalization relies on a dense sampling of the pose space so the limiting factor on the sampling density is computation time. In this evaluation, the number of samples has been chosen such that the computation time is well under a second. The resolution of rotation space has been prioritized, so the rotational space has been sampled using 3 HEALPix recursions ($72 * 8^3 = 36864$ samples) while the spatial dimensions perpendicular to the camera axis are sampled in an 11x11 grid around an estimated object position (121 samples covering a 10mm × 10mm area) for a total of 4.46 million samples. The computation of all samples takes approximately 100 ms on a GeForce GTX 1080 Ti

GPU. The marginal rotation distribution is then computed by numerically integrating over grid positions: $\mathrm{p}(R_i|I) = \int \mathrm{p}(R_i,t|I)dt \approx \left(\sum_{t \in t_{\text{grid}}} \hat{\mathrm{P}}(R_i,t|I)\right)/V_{\text{grid}}$

As stated in Sec. 3.2 it is assumed that an object detector provides a bounding box around the object. In the evaluation, the bounding box is chosen to be 20% larger than a tight crop, and uncertainty on estimation is simulated by using the ground truth bounding box with a random scaling of ±5% and a random translation computed such that the object remains fully inside the crop. Furthermore, it is assumed that an estimate of the object position is provided by the object detector, which is simulated by using the ground truth object position with added uniform noise of $\pm 10mm$ on all three dimensions. The spatial grid in the normalization is centered around this estimate.

The implementation is evaluated on the test part of the BOP version of the YCB-V and T-LESS datasets. The rotation distribution across the SO(3) grid is estimated and the log-likelihood of the ground truth rotation is chosen as the closest sample in the grid. The mean loglikelihood score (meanLL) for a single object is then computed by averaging across all images of the object and the dataset score is computed by averaging over the meanLL scores for objects in the dataset. The results of the evaluation on the YCBV and T-less datasets are compared to the results published in [14] and [12] respectively.

The evaluation also presents a qualitative evaluation of our method's interpretability in the form of heatmaps and orientation distributions for three representative objects. The visualization of the rotation distribution is inspired by [12]. For the visualization, an orientation is parameterized by the direction of the object's canonical x-axis expressed as longitude and latitude (a,b), and the tilt c around the x-axis. The distribution is then visualized by plotting (a,b) using a Mollweide projection, indicating c with color, and indicating the probability mass with the alpha value: $\alpha_i = P_i/P_{\max}$. The ground truth pose is indicated with an x. The visualization is created using the ground truth crop and object position, and 5 HEALPix recursions.

To investigate the impact of sampling density on the meanLL score, the score has been computed for increasing numbers of recursions of the HEALPix method, using the ground truth crop and object position. The number of samples also determines the theoretical maximum of the log-likelihood, with the limit being when a single sample has a probability mass of one, which corresponds to a log-likelihood of $V_{\text{grid}}^{-1} = M/\pi^2$.

5 Results

The evaluation in [14] evaluated 9 different approaches for pose distribution estimation. Each approach was implemented twice using either DenseFusion or PoseCNN as the basis of the implementation, which resulted in a comparison of 18 different methods. Our analysis of the reported results reveals that four of the methods can account for all but one of the per-object best results. The condensed summary shown in Table 1 shows that no single method was able to perform well on all objects. The results show that our method is superior for most objects. The relatively low meanLL for the wood block and bowl can be attributed to the many discrete and continuous symmetries respectively. In such cases, there will be many orientations with a low probability that should have been zero. These regions are attenuated due to the approximation of Eq. 5, leading to dispersed probability distributions. It is important to note that unlike method I, II, and IV, our method doesn't have a meanLL score less than -2.29 which correspond to a uniform distribution on SO(3).

Deng[1]	5.3
Gilitschenski[1]	6.9
Prokudin[1]	8.8
IPDF[1]	9.8
Ki-Pode (ours)	3.3

Table 2: MeanLL score on the T-LESS dataset ([1]Results published in [12]).

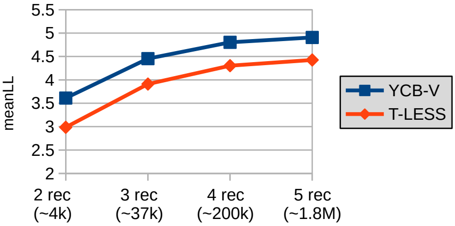

Figure 2: The impact of sampling density on the meanLL score. The x-axis shows the number of healpix recursions used to generate the equivolumetric grid on SO(3), and the corresponding number of samples. The theoretical upper bound on the log likelihood for recursions 2-5 is respectively 6.1, 8.2, 10.3, and 12.4.

The evaluation of different state-of-the-art methods on the T-LESS dataset published in [12] does not contain per object evaluations, so only the meanLL score over all objects is computed. The scores from [12] and our method is shown in Table 2. The T-LESS dataset contains many objects with near rotational symmetry, which for our method, as discussed for the bowl in YCB-V, leads to relatively modest performance, presumably because we only model keypoint marginals. Our method is fundamentally different from previous work and can be considered an early work towards interpretable and translational equivariant representations for pose distributions with great potential for further improvements. For instance, an efficient auto-regressive model using this formulation would be able to represent the joint keypoint distributions.

To qualitatively evaluate the interpretability of our method, Fig. 1 shows image crops, the logarithm of heatmaps, and rotation distribution for three representative objects. The first row shows an object for which most keypoints are well localized, leading to a single well-defined orientation. The second row shows an object with a discrete 180-degree rotation symmetry which in the heatmaps appears as two peaks. The third row shows an object with near rotational symmetry. It is clear from the heatmaps, that the U-Net is able to locate the keypoints as lying on ellipses, but cannot detect that there is only near rotational symmetry. Furthermore, the ellipses are dispersed which leads to uncertainty of the direction of the symmetry axis. The resulting orientation distribution appears as a continuous region on SO(3) as shown in the orientation visualization.

The impact of sample density on the meanLL score is shown in figure 2. The analysis indicates that increasing sample density improves the score with diminishing returns as the number of samples increases. The computation time scales linearly with the number of samples, which increases by a factor of 8 for each HEALPix recursion. As discussed in Sec. 4, using 3 recursions and an 11×11 spatial grid takes approximately 100ms, so a rough estimate of the computation time with 4 or 5 recursions is 0.8s or 6.4s respectively. Ultimately, the choice of how many recursions to use will depend on the available computational budget.

6 Conclusion

We have proposed a novel method for estimating a probability density function over 6D object poses. The distribution is formulated implicitly using keypoints as an intermediary object representation which ensures a high expressiveness of the distribution as well as a high level of interpretability of the estimates. The proposed method approximates the probability distribution from marginal distributions over object keypoints and is based on conservative approximations which leads to high reliability of the results. The method has been evaluated on the YCB-V and T-LESS datasets and has been confirmed to perform reliably on all objects.

Acknowledgements

This project was funded in part by Innovation Fund Denmark through the project "MADE FAST - Flexible, Agile, Sustainable manufacturing enabled by Talented employees"

References

[1] Alexander Buslaev, Vladimir I. Iglovikov, Eugene Khvedchenya, Alex Parinov, Mikhail Druzhinin, and Alexandr A. Kalinin. Albumentations: Fast and flexible image augmentations. *Information*, 11(2), 2020. ISSN 2078-2489. doi: 10.3390/info11020125. URL https://www.mdpi.com/2078-2489/11/2/125.

[2] Haowen Deng, Mai Bui, Nassir Navab, Leonidas Guibas, Slobodan Ilic, and Tolga Birdal. Deep bingham networks: Dealing with uncertainty and ambiguity in pose estimation. *International Journal of Computer Vision*, pages 1–28, 2022.

[3] Xinke Deng, Arsalan Mousavian, Yu Xiang, Fei Xia, Timothy Bretl, and Dieter Fox. Poserbpf: A rao–blackwellized particle filter for 6-d object pose tracking. *IEEE Transactions on Robotics*, 37(5):1328–1342, 2021.

[4] Maximilian Denninger, Martin Sundermeyer, Dominik Winkelbauer, Youssef Zidan, Dmitry Olefir, Mohamad Elbadrawy, Ahsan Lodhi, and Harinandan Katam. Blenderproc. *arXiv preprint arXiv:1911.01911*, 2019.

[5] Marek Franaszek and Geraldine S Cheok. Propagation of orientation uncertainty of 3d rigid object to its points. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 2183–2191, 2017.

[6] Igor Gilitschenski, Roshni Sahoo, Wilko Schwarting, Alexander Amini, Sertac Karaman, and Daniela Rus. Deep orientation uncertainty learning based on a bingham loss. In *International Conference on Learning Representations*, 2019.

[7] Krzysztof M Gorski, Eric Hivon, Anthony J Banday, Benjamin D Wandelt, Frode K Hansen, Mstvos Reinecke, and Matthia Bartelmann. Healpix: A framework for high-resolution discretization and fast analysis of data distributed on the sphere. *The Astrophysical Journal*, 622(2):759, 2005.

[8] Tomáš Hodaň, Martin Sundermeyer, Bertram Drost, Yann Labbé, Eric Brachmann, Frank Michel, Carsten Rother, and Jiří Matas. BOP challenge 2020 on 6D object localization. *European Conference on Computer Vision Workshops (ECCVW)*, 2020.

[9] Fabian Manhardt, Diego Martin Arroyo, Christian Rupprecht, Benjamin Busam, Tolga Birdal, Nassir Navab, and Federico Tombari. Explaining the ambiguity of object detection and 6d pose from visual data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6841–6850, 2019.

[10] Zoltán Csaba Márton, Serkan Türker, Christian Rink, Manuel Brucker, Simon Kriegel, Tim Bodenmüller, and Sebastian Riedel. Improving object orientation estimates by considering multiple viewpoints. *Autonomous Robots*, 42(2):423–442, 2018.

[11] David Mohlin, Josephine Sullivan, and Gérald Bianchi. Probabilistic orientation estimation with matrix fisher distributions. *Advances in Neural Information Processing Systems*, 33:4884–4893, 2020.

[12] Kieran A Murphy, Carlos Esteves, Varun Jampani, Srikumar Ramalingam, and Ameesh Makadia. Implicit-pdf: Non-parametric representation of probability distributions on the rotation manifold. In *Proceedings of the 38th International Conference on Machine Learning*, pages 7882–7893. PMLR, 2021.

[13] Aiden Nibali, Zhen He, Stuart Morgan, and Luke Prendergast. Numerical coordinate regression with convolutional neural networks. *arXiv preprint arXiv:1801.07372*, 2018.

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

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

[16] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241, Cham, 2015. Springer International Publishing. ISBN 978-3-319-24574-4.

[17] Guanya Shi, Yifeng Zhu, Jonathan Tremblay, Stan Birchfield, Fabio Ramos, Animashree Anandkumar, and Yuke Zhu. Fast uncertainty quantification for deep object pose estimation. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5200–5207. IEEE, 2021.

[18] Hao Su, Charles R Qi, Yangyan Li, and Leonidas J Guibas. Render for cnn: Viewpoint estimation in images using cnns trained with rendered 3d model views. In *Proceedings of the IEEE international conference on computer vision*, pages 2686–2694, 2015.

[19] Martin Sundermeyer, Zoltan-Csaba Marton, Maximilian Durner, Manuel Brucker, and Rudolph Triebel. Implicit 3d orientation learning for 6d object detection from rgb images. In *Proceedings of the european conference on computer vision (ECCV)*, pages 699–715, 2018.

[20] Naoto Usuyama and Karanbir Chahal. pytorch-unet, github repository, Commit: 2020/08/21. URL `https://github.com/usuyama/pytorch-unet`.

[21] Anna Yershova, Swati Jain, Steven M Lavalle, and Julie C Mitchell. Generating uniform incremental grids on so (3) using the hopf fibration. *The International journal of robotics research*, 29(7):801–812, 2010.