6th IEEE International Conference on Robotic Computing (IRC), Naples, Italy, December 2022.

Learning Implicit Probability Distribution Functions for Symmetric Orientation Estimation from RGB Images Without Pose Labels

Arul Selvam Periyasamy*
Autonomous Intelligent Systems
University of Bonn
Bonn, Germany
Email: periyasa@ais.uni-bonn.de

Luis Denninger*
Autonomous Intelligent Systems
University of Bonn
Bonn, Germany
Email: l_denninger@uni-bonn.de

Sven Behnke
Autonomous Intelligent Systems
University of Bonn
Bonn, Germany
Email: behnke@cs.uni-bonn.de

***Abstract*—Object pose estimation is a necessary prerequisite for autonomous robotic manipulation, but the presence of symmetry increases the complexity of the pose estimation task. Existing methods for object pose estimation output a single 6D pose. Thus, they lack the ability to reason about symmetries. Lately, modeling object orientation as a non-parametric probability distribution on the SO(3) manifold by neural networks has shown impressive results. However, acquiring large-scale datasets to train pose estimation models remains a bottleneck. To address this limitation, we introduce an automatic pose labeling scheme. Given RGB-D images without object pose annotations and 3D object models, we design a two-stage pipeline consisting of point cloud registration and render-and-compare validation to generate multiple symmetrical pseudo-ground-truth pose labels for each image. Using the generated pose labels, we train an ImplicitPDF model to estimate the likelihood of an orientation hypothesis given an RGB image. An efficient hierarchical sampling of the SO(3) manifold enables tractable generation of the complete set of symmetries at multiple resolutions. During inference, the most likely orientation of the target object is estimated using gradient ascent. We evaluate the proposed automatic pose labeling scheme and the ImplicitPDF model on a photorealistic dataset and the T-Less dataset, demonstrating the advantages of the proposed method.**

I. Introduction

6D object pose estimation is the task of predicting the translation $\mathbf{t} \in \mathbb{R}^3$ and the orientation $\mathbf{R} \in \mathbf{SO}(3)$ of an object in the sensor coordinate frame. It is a necessary prerequisite for autonomous robotic manipulation, industrial bin picking, as well as virtual and augmented reality. With the advent of convolutional neural networks (CNNs), significant progress has been made in object estimation from RGB and RGB-D images. We use the notation RGB-(D) to denote either RGB or RGB-D images. Despite the improvements, large-scale object pose estimation remains challenging. The challenges, particularly in orientation estimation, are compounded by object symmetries. Objects in our daily life and in industrial setups exhibit symmetries due to which it is impossible to estimate a single 6D pose only. Ambiguities due to symmetries present hindrance in learning visual representations for pose estimation. Formally, ambiguities occur when an object $\mathcal{O}$ appears similar under at least two different poses $\mathbf{P}_i$ and $\mathbf{P}_j$, i.e., we obtain the same image $\mathcal{I}$ when object $\mathcal{O}$ is in pose $\mathbf{P}_i$ and $\mathbf{P}_j$:

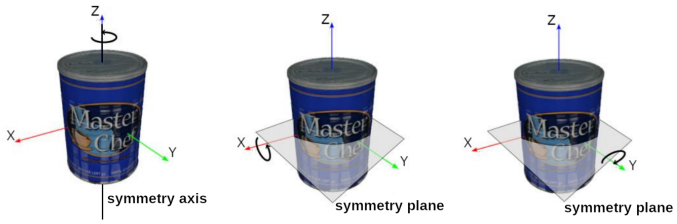

Fig. 1. *Can* object symmetries. Due to the presence of visual-symmetry-breaking texture, the *can* object exhibits only geometric symmetries, namely a continuous rotational symmetry along the z axis and a discrete flip symmetry along the x and y axes.

$$\mathcal{I}\left(\mathcal{O}, \mathbf{P}_i\right) = \mathcal{I}\left(\mathcal{O}, \mathbf{P}_j\right). \tag{1}$$

Symmetries can be classified into visual symmetries and geometric symmetries. Visual symmetries, as defined in Eq. (1), arise due to the lack of distinctive visual features, whereas even in the presence of symmetry-breaking visual features, objects may exhibit symmetries in terms of their geometry. Knowing the object axes and type of symmetries, we can describe symmetries—despite the presence of textures—with respect to the object's geometry as discrete or continuous rotations around the object axes, as shown in Fig. 1. Formally, an object $\mathcal{O}$ consisting of n 3D points $\mathbf{x}$ can be considered to exhibit geometric symmetries when there are at least two poses $\mathbf{P}_i$ and $\mathbf{P}_j$ that have a small mean closest point distance:

$$\frac{1}{n} \sum_{\mathbf{x}_1 \in \mathcal{O}} \min_{\mathbf{x}_2 \in \mathcal{O}} ||\mathbf{P}_i \mathbf{x}_1 - \mathbf{P}_j \mathbf{x}_2|| \approx 0. \tag{2}$$

Enumerating all sources of symmetries is not tractable, making an approach as shown in Fig. 1 to describe arbitrary object symmetries not scalable.

Following the terminology introduced by Brégier *et al.* [1], we define *proper symmetries* $\mathcal{M}$ as the group of poses that exhibit geometric symmetries:

$$\begin{aligned} \mathcal{M} = \{&\mathbf{m} \in \mathbf{SE}(3) \text{ such that} \\ &\forall \mathbf{P} \in \mathbf{SE}(3), \frac{1}{n} \sum_{\mathbf{x}_1 \in \mathcal{O}} \min_{\mathbf{x}_2 \in \mathcal{O}} ||\mathbf{P}_i \mathbf{x}_1 - \mathbf{m} \cdot \mathbf{P}_j \mathbf{x}_2|| \approx 0.\}, \end{aligned} \tag{3}$$

with $\mathbf{m}$ being the transformations rotating the object around

* Equal contribution

its symmetry axes in discrete steps.

The methods for pose estimation for symmetric objects can be classified into two families.

The first family of methods estimates a single valid pose $\mathbf{p} \in \mathbf{SE}(3)$ corresponding to the RGB-(D) image. One major advantage of this approach is that the CNNs architectures remain the same for symmetric and non-symmetric objects. In both cases, given an RGB-(D) image, the network generates a single 6D pose prediction. The only difference lies in the loss function used to train the model. Thus, in terms of the neural network architecture, training scheme, and inference, both symmetric and non-symmetric objects are treated the same.

The second family of methods estimates the complete set of *proper symmetries* $\mathcal{M}$. Modeling symmetries explicitly provides benefits that are twofold: i) facilitate models in learning better visual representations and ii) better integration with downstream tasks.

Our work is based on the implicit probability distribution (ImplicitPDF) models for rotation manifolds introduced by Murphy *et al.* [2]. Given an RGB image and a pose hypothesis, we train a CNN to estimate the likelihood of the pose hypothesis given the RGB image. The novelty of our approach is that we do not need explicit ground-truth pose annotations to train the ImplicitPDF CNN model, eliminating the bottleneck of acquiring large-scale ground-truth annotated dataset.

Given RGB-D images of symmetric objects without pose annotations, we propose an automatic pose labeling scheme and train the ImplicitPDF model using the generated pseudo ground-truth. The model trained with the pseudo ground-truth is able to express the complete set of *proper symmetries* $\mathcal{M}$ without prior knowledge about object symmetries.

II. Related Work

The state-of-the-art methods for predicting a single 6D pose are trained using ShapeMatch-Loss [3] for symmetric objects [4]–[7]. Employing ShapeMatch-Loss implicitly selects one pose closest to the current pose prediction from the *proper symmetry* set as ground-truth. While ShapeMatch-Loss does not need explicit definition of symmetry, during training, the ground-truth pose depends on the current pose prediction and this variability in ground-truth pose might hamper the models' learning ability. In contrast, Pitteri *et al.* [8] and Periyasamy *et al.* [9] mapped the symmetrical rotations to a single "canonical" rotation. One disadvantage of these methods is the requirement of explicit definition of object symmetry. Esteves *et al.* [10] and Saxena *et al.* [11] proposed methods to learn features equivariant to specific symmetry classes. In addition to learning pose estimation, Rad and Lepetit [12] added an auxiliary task to classify the type of symmetry an object exhibits. The authors argued that the auxiliary task helped the model to learn additional properties of the object's symmetry, which, in turn, benefit 6D pose estimation. While the auxiliary task helped improving the single pose prediction accuracy, the formulation of the auxiliary task as a classification task limits its scope in modeling *proper symmetries*. Corona *et al.* [13] sidestepped the problem of estimating 6D pose and modeled pose estimation as a task of image comparison. They trained a model to estimate a similarity score of two RGB images and used the similarity score to select the image from a codebook of images that best matches the test image during inference. The pose corresponding to the matched image is considered as the pose of the target object in the test image. In case of symmetric objects, multiple images from the codebook have a high similarity score. Since the inference involves comparing against a large-size codebook of images, the inference time requirement is high. Sundermeyer *et al.* [14] addressed this issue by using *Augmented Autoencoder*, a variant of Denoising Autoencoder, to learn a low-dimensional latent space representation for images and uses the latent space for image comparison. Despite the speed-up achieved in the codebook comparison, discretization of $\mathbf{SO}(3)$ is still needed to make the model real-time capable. Manhardt *et al.* [15] trained their model to generate a set of predictions given an RGB image with an intent to cover multiple possible poses. In their experiments, they set the number of pose predictions to five. While their model can predict up to five correct poses in the presence of visual symmetry, it is neither enough to cover the complete *proper symmetries* nor does it reveal any information about the type of symmetry. Deng *et al.* [16] and Gilitschenski *et al.* [17] modeled multiple pose hypotheses as Bingham distributions and trained a CNN model to estimate the distribution parameters given an RGB-D input. Mohlin *et al.* [18] used Fisher distributions to model multiple pose hypotheses. These methods suffer under the complexity that arises in estimating the normalization constant of the distributions. Okorn *et al.* [19] predicted a discrete histogram distribution of pose hypotheses by learning to predict the likelihood using comparison against a dictionary of images. Our approach to model symmetries is inspired by ImplicitPDFs for rotational manifold introduced by Murphy *et al.* [2], which we discuss in Section III-A in detail. In contrast to Murphy *et al.* [2], we propose an automatic pose labeling scheme to generate multiple ground-truth annotations for each training image to train our ImplicitPDF model.

III. Method

A. Implicit Probability Distribution for Object Orientation

Inspired by the success of Neural Fields in modeling shape, scene, and rotation manifold [2], [20]–[22], we model object symmetries as a conditional probability distribution of the likelihood $\mathcal{P}(\mathbf{P} \mid \mathcal{I})$ of the pose hypothesis $\mathbf{P}$ given an image $\mathcal{I}$ implicitly using a neural network. To this end, we train a neural network $\mathcal{F}$ to predict the unnormalized joint log probability $\mathcal{F}(\mathbf{P}, \mathcal{I})$ of the pose hypothesis $\mathbf{P}$ and image $\mathcal{I}$ as shown in Fig. 2. Let α be the normalization constant such that

$$\mathcal{P}(\mathbf{P}, \mathcal{I}) = \alpha \exp(\mathcal{F}(\mathbf{P}, \mathcal{I})). \tag{4}$$

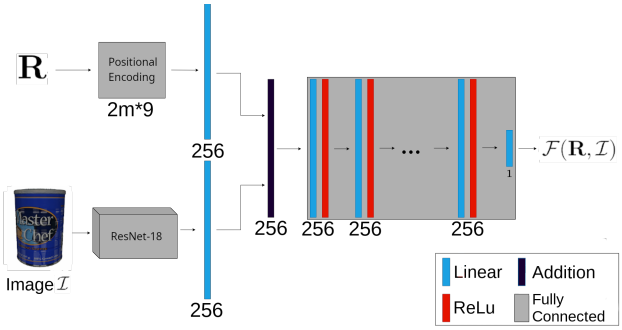

Fig. 2. Learning ImplicitPDF. Given an image $\mathbf{I}$ and an orientation hypothesis $\mathbf{R}$, the CNN model is trained to generate the unnormalized joint log probability of the orientation hypothesis and the image.

Using the product rule,

$$\mathcal{P}(\mathbf{P} \,|\, \mathcal{I}) = \frac{\mathcal{P}(\mathbf{P}, \mathcal{I})}{\mathcal{P}(I)}, \tag{5}$$

where

$$\mathcal{P}(I) = \int_{\mathbf{P} \in \mathbf{SE}(3)} \mathcal{P}(\mathbf{P}, \mathcal{I}) d\mathcal{I}. \tag{6}$$

For simplicity, we consider only the object orientation $\mathbf{R} \in \mathbf{SO}(3)$ instead of the 6D pose $\in \mathbf{SE}(3)$. To make computing marginal probabilities tractable, we replace the continuous integral in Eq. (6) with a discrete summation over a equivolumetric partitioning of $\mathbf{SO}(3)$ with N partitions of volume $V = \pi^2/N$, and cancel out the normalization constant α:

$$\mathcal{P}(\mathbf{R} \,|\, \mathcal{I}) = \frac{1}{V} \frac{\exp(\mathcal{F}(\mathbf{R}, \mathcal{I}))}{\sum_i^N \exp(\mathcal{F}(\mathbf{R}_i, \mathcal{I}))}. \tag{7}$$

For a detailed derivation of Eq. (7), we refer to [2].

1) Training: We train our model to minimize the negative log-likelihood of the ground-truth orientation $\mathbf{R}_{GT}$:

$$\mathcal{L}(\mathcal{I}, \mathbf{R}_{GT}) = -\log(\mathcal{P}(\mathbf{R}_{GT} | \mathcal{I})). \tag{8}$$

Following Eq. (7), we approximate the computation of the distribution $\mathcal{P}(\mathbf{R}_i | \mathcal{I})$ using $\mathbf{R}_i \in \{\mathbf{R}^0\}$, an equivolumetric grid covering $\mathbf{SO}(3)$ as in Section III-A3. The orientation hypothesis $\mathbf{R}$ given to the model as input is represented using *positional encoding* [23].

2) Inference: During inference, given an image $\mathcal{I}$, we predict the (single) most plausible orientation $\mathbf{R}_I^*$ using gradient ascent starting from a set of initial hypotheses $\{\mathbf{R}^0\}$:

$$\mathbf{R}_I^* = \underset{\mathbf{R} \in \mathbf{SO}(3)}{\arg\max}\, \mathcal{F}(\mathbf{R}, \mathcal{I}). \tag{9}$$

To generate the full distribution, we evaluate $\mathcal{P}(\mathbf{R}j \,|\, \mathcal{I}) : \mathbf{R}_j \in \{\mathbf{R}^n\}$ sampled equivolumetrically over $\mathbf{SO}(3)$.

3) Equivolumetric Sampling and Visualization of $\mathbf{SO}(3)$*:* We follow the equivolumetric sampling of rotation manifold approach proposed by Murphy *et al.* [2] to generate $\{\mathbf{R}^0\}$ and $\{\mathbf{R}^n\}$ to cover $\mathbf{SO}(3)$ at different resolutions. Using the HEALPix algorithm [24] as a starting step, we generate equal area grids on the 2-sphere and iteratively use Hopf fibration [25] to follow a great circle through each point on the surface of a 2-sphere to cover $\mathbf{SO}(3)$. We also use the visualization method proposed by Murphy *et al.* [2] to visualize distributions of object orientations on the $\mathbf{SO}(3)$ manifold. Rotation matrices in $\mathbf{SO}(3)$ have three degrees of freedom—two of the degrees of freedom are represented as a 2-sphere and projected on to a plane using Mollweide projection. The third degree of freedom is represented using Hopf fibration by a great circle of points to each point on the 2-sphere. The location of a point on the great circle is represented using a color wheel as shown in Fig. 3. The number of samples generated in iteration S_i is given by $72 \cdot 8^{S_i}$.

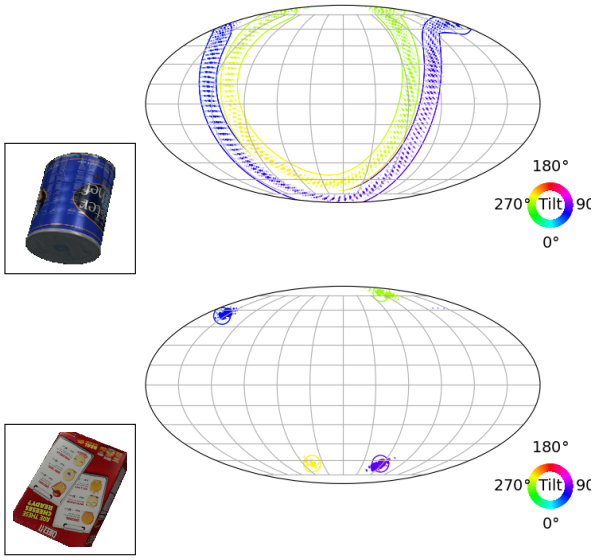

Fig. 3. Visualization of the ground-truth and the orientations predicted by the ImplicitPDF model for *can* and *box* objects. Given an RGB image and an orientation hypothesis, the ImplicitPDF model estimates the likelihood. The continuous lines and circles represent the ground-truth symmetries and the dots represent the orientation hypotheses with a high estimated likelihood. Two of three degrees of freedom of the SO(3) manifold are represented as a 2-sphere and projected on to a plane using Mollweide projection. The third degree of freedom is represented using a point on the color wheel.

4) Evaluation Metrics: To evaluate the performance of the proposed method, we use the *log-likelihood* (LLH) and the *mean absolute angular error* (MAAD) metrics. Given a set of ground-truth annotations $\mathbf{R}_{GT}$, the LLH metric measures the likelihood of the ground-truth orientations:

$$\text{LLH}(\mathbf{R}) = \mathbb{E}_{I \sim \mathcal{P}(I)}\, \mathbb{E}_{\mathbf{R} \sim \mathcal{P}_{\mathcal{GT}}(\mathbf{R}|\mathcal{I})} \log(\mathcal{P}(\mathbf{R}|\mathcal{I})).$$

To compute log-likelihood, we do not need the complete set of *proper symmetries*. The standard *mean absolute angular deviation* (MAAD) is defined as:

$$\text{MAAD}(\mathbf{R}) = \mathbb{E}_{\mathbf{R} \sim \mathcal{P}(\mathbf{R}|\mathcal{I})} \left[\min_{\mathbf{R}' \in \mathbf{R}_{GT}} d(R, R')\right],$$

where d is the geodesic distance between rotations.

Furthermore, we report Recall MAAD as a measure of recall. We extract a set of orientations $\{\hat{\mathbf{R}}\}$ with a predicted probability threshold of 1e-3. For each orientation in $\{\mathbf{R}_{GT}\}$,

we find the closest orientation in $\{\hat{\mathbf{R}}\}$ in terms of the geodesic distance and report the mean of the shortest angular distance over the set $\{\mathbf{R}_{GT}\}$. In the case of continuous symmetries, we discretize the symmetries into 200 orientations for computing the Recall MAAD metric.

B. Learning Without Ground-Truth Annotations

Murphy *et al.* [2] introduced the SYMSOL I and SYMSOL II datasets to benchmark symmetry learning methods. The datasets consist of renderings of platonic solids (tetrahedron, cube, icosahedron), cone, and cylinder and corresponding ground-truth symmetries. In real-world applications, acquiring ground-truth symmetry annotations is prohibitively expensive. To address this issue, we propose a two-stage automatic pose labeling scheme as illustrated in Fig. 4.

Given an RGB-D image of a scene and the 3D mesh of the target object, we start with unprojecting the depth map into 3D to generate the observed point cloud C_{obs}. We transform the model point cloud C_{model} according to a random initial pose P_0 and perform global registration of the transformed model point cloud against the observed point cloud to generate a pose hypothesis P_1. We employ the *Fast Global Registration* algorithm [26] in conjunction with *Fast Point Feature Histogram* (FPFH) feature [27] to perform global registration. We refine P_1 using the *Iterative Closest Point* (ICP) algorithm to generate $\hat{P}$. The results of the different stages of the pipeline are visualized in Fig. 5. Transforming C_{model} to a random initial pose P_0 at the beginning ensures that we generate a different $\hat{P}$ every time we run the process for an RGB-D image. This way, we generate a set of pseudo ground-truth pose labels for each image in the training set. The variability in the set of generated pseudo ground-truth pose labels for an image is vital for the model in learning the complete set of *proper symmetries*. Due to self-occlusion, an object is only partially visible in an RGB-D image. Without the knowledge of camera view direction, registering the complete model point cloud C_{model} against the partial observed point cloud C_{obs} might result in bad registrations and it is not possible to detect the bad registrations based on the standard ℓ_2 distance metric. To address this issue, we utilize the render-and-compare framework. We render the depth map according to $\hat{P}$ and compare it pixel-wise with the observed depth map. In the ideal scenario, the render-and-compare difference should be close to zero. To generate one pseudo ground-truth label, we repeat the process multiple times and select the $\hat{P}$ with the smallest comparison score.

C. Dataset

To evaluate the proposed method, we generate a photo-realistic dataset using the Isaac GYM framework [28]. The dataset consists of three objects—*can*, *box*, and *bowl* (shown in Fig. 6)—placed in randomly-sampled physically-plausible poses on a tabletop. Each frame in the dataset consists of an RGB-D image, 6D object pose (single pose used to render the image) and segmentation ground-truth. Moreover, to evaluate the impact of the object texture over the model's ability to learn symmetries, we generate two versions of the dataset: using a uniform colored texture and using the original material texture. We call these datasets *Uniform* and *Texture*, respectively. Both datasets consist of 20K images. 20% of the samples in each dataset are used for validation and the rest are used for training.

IV. Experiments

TABLE I
Results of models trained on different ground truths.

	GT	Without Occlusion			With Occlusion		
		LLH ↑	MAAD [°] ↓	Recall MAAD [°] ↓	LLH ↑	MAAD [°] ↓	Recall MAAD [°] ↓
can	Single	6.325	2.922	115.375	4.184	5.225	76.354
	Analytical	3.276	4.413	2.077	3.38	5.039	2.079
	Pseudo	2.359	2.457	2.09	2.712	5.043	2.003
box	Single	7.096	2.651	72.586	6.716	5.386	57.292
	Analytical	3.656	3.444	1.978	3.436	4.955	1.955
	Pseudo	4.452	7.481	2.31	4.544	8.845	2.353
bowl	Single	6.713	4.157	121.233	6.467	5.157	119.141
	Analytical	4.39	8.534	2.839	3.835	9.266	2.223
	Pseudo	4.544	8.845	2.353	3.835	9.266	2.163
Pseudo Avg.		3.785	6.261	2.178	3.697	7.715	2.173

↑ indicates higher value better, whereas ↓ indicates lower value better.

We train our model for 50 epochs with 200 iterations per epoch. Each iteration consists of a batch of 64 images. We assume that the object bounding box and the object segmentation mask are available to us. Using the bounding box, we extract a crop of 560×560 and resize it to the standard ResNet input size 224×224. Additionally, we mask the background pixels using the segmentation mask. To compute $\mathcal{P}(\mathbf{R_{GT}} \,|\, \mathcal{I})$ we use a grid $\{\mathbf{R^0}\}$ of cardinality 4,608 (S_i=2). During testing, we compute the evaluation metrics using $\{\mathbf{R^n}\}$ of cardinality 294,912 (S_i=4).

Our model learns to predict the orientation on both Uniform and Texture datasets. In Fig. 7, we show the training loss and the log-likelihood metric on the validation set during the training process. We early stop the training when the log-likelihood metric starts to stagnate. Qualitative samples of the predicted orientation distribution are shown in Fig. 8. From the visualizations, we observe that the model learns to predict the complete set of *proper symmetries*. In Table I, we report the validation scores. Overall, in the absence of occlusions, our model achieves a MAAD score of ∼6.2° and a Recall MAAD score of ∼2.2°. A lower MAAD indicates high accuracy of the orientation predictions and a lower Recall MAAD shows that the model learns the complete set of *proper symmetries*.

A. Occlusion

Occlusion increases the complexity of computer vision problems. Pose estimation, in particular, is heavily impacted by the presence of occlusion. To make our model robust against occlusion, we train our model by masking out random crops in the input images. We augment 80% of images in each training

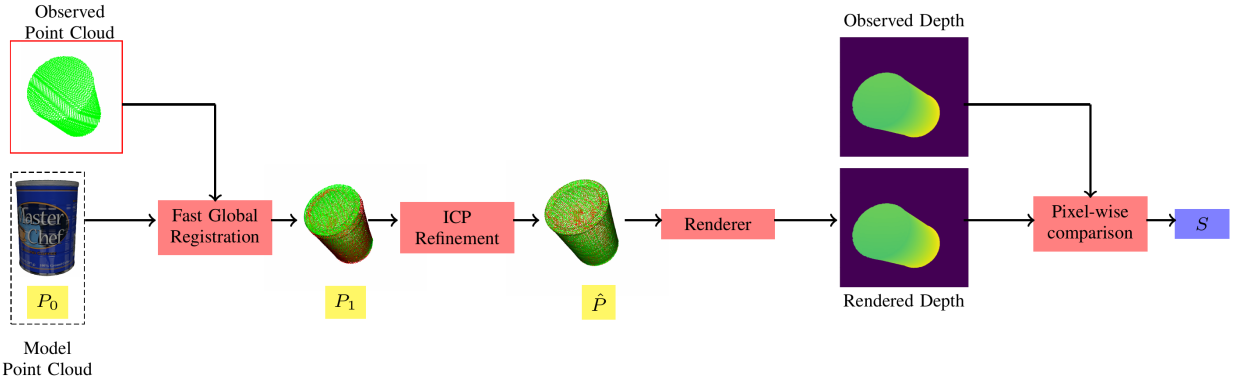

Fig. 4. Pseudo ground-truth generation process. Given an RGB-D image without pose annotation, we register the model point cloud in a random pose P_0 against the observed point cloud using the Fast Global Registration algorithm to generate hypothesis P_1, which is further refined using the ICP algorithm utilizing the depth information to generate $\hat{P}$. We then render the model according to $\hat{P}$ to generate the depth image. The rendered depth image is compared with the observed depth image pixel-wise to compute the comparison score S. For a given RGB-D image we run this process multiple times and select the $\hat{P}$ with the smallest S.

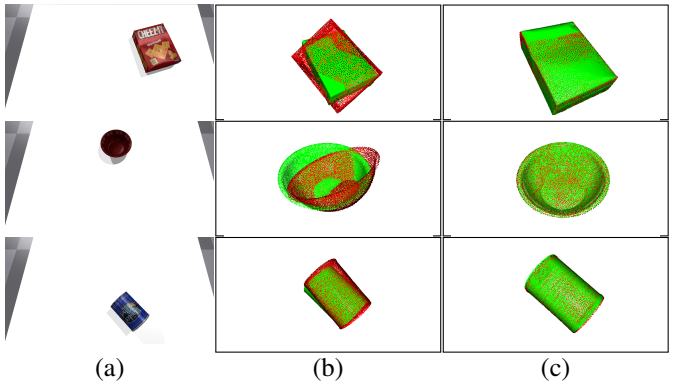

Fig. 5. Visualization of pseudo ground-truth generation. (a) RGB image of the scene. (b) Pose generated by the Fast Global Registration algorithm. (c) Final pseudo ground-truth pose generated by ICP optimization. In (b) and (c), the point cloud in ground-truth pose are visualized in green and the point cloud in generated pseudo ground-truth poses are visualized in red.

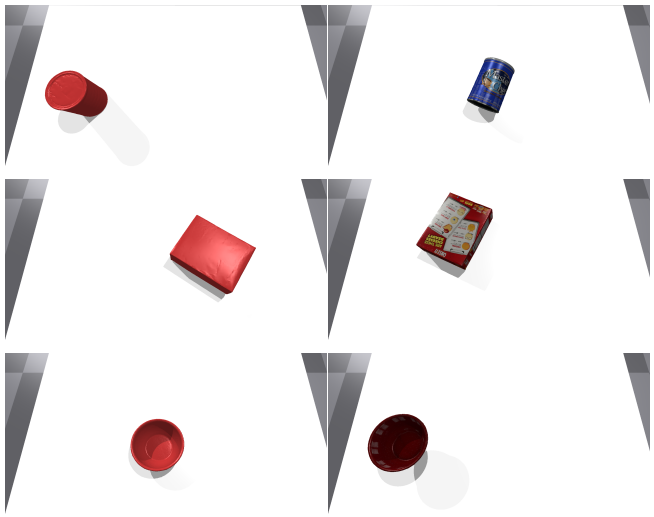

Fig. 6. Exemplar RGB images from the Uniform and Texture dataset. First column: Uniform dataset. Second column: Texture dataset.

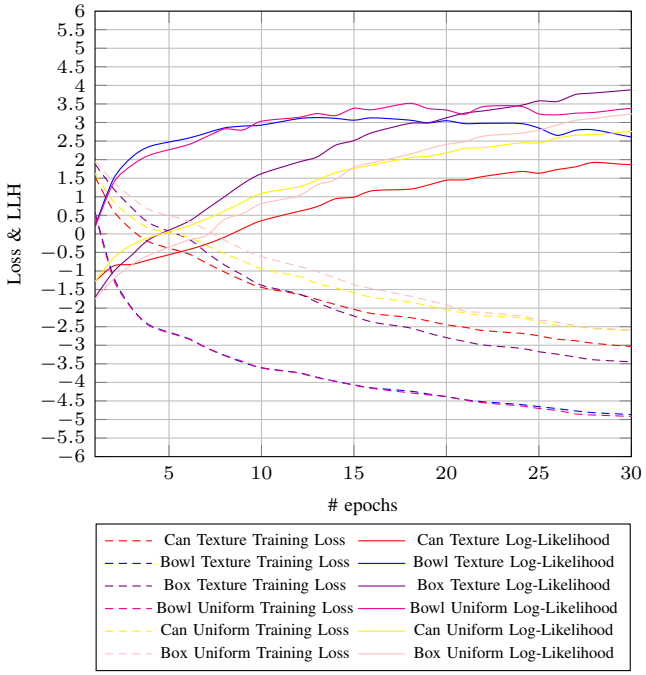

Fig. 7. Training loss and log-likelihood on the validation set during the training process on the Uniform and Texture dataset. We early stop the training after 30 epochs when the log-likelihood starts stagnating.

batch randomly. Between 10% and 50% of the image portions are occluded. In Fig. 10, we present qualitative samples of the orientation predicted by our model in the presence of occlusion. Despite the presence of occlusion, our model learns the orientation distribution well on both Texture and Uniform dataset. Our model performs only slightly worse compared to the occlusion-free model. Quantitatively, in the presence of occlusion, our model achieves a MAAD score of $\sim$7.7° and a Recall MAAD score of $\sim$2.2°. Interestingly, in terms of the Recall MAAD metric, the model trained using the single ground-truth performs significantly better than in the case of no occlusions. This can be attributed to the uncertainty in the

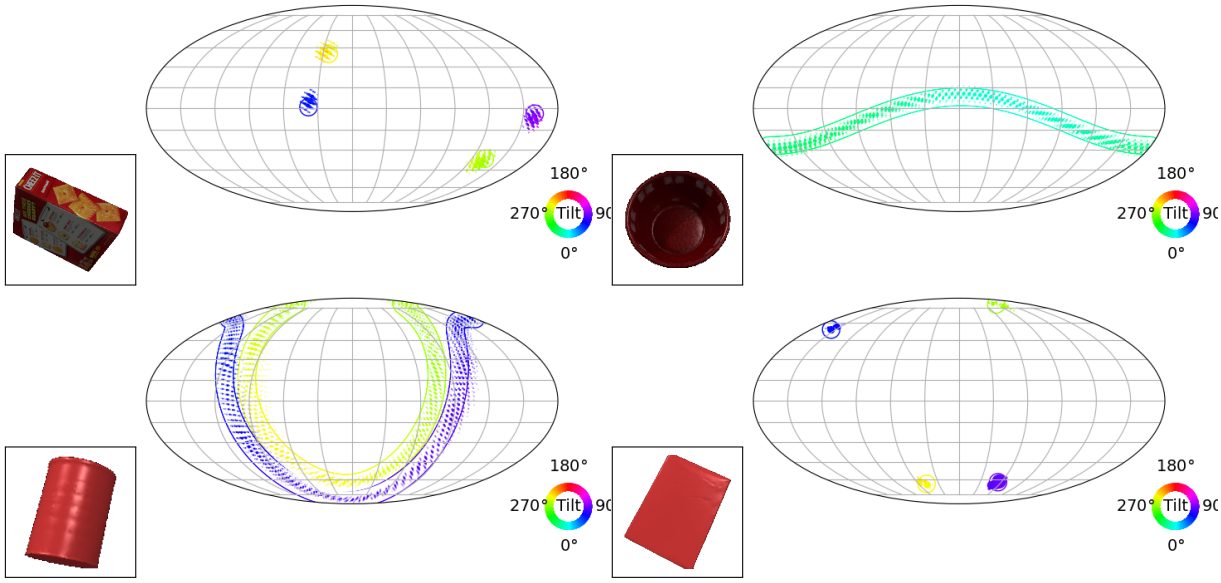

Fig. 8. Orientation distributions predicted by our model. Top: Texture dataset. Bottom: Uniform dataset. The continuous lines and the circles represent the ground-truth symmetries and the dots represent the orientation hypotheses with a high estimated likelihood. The visualizations are generated using 294,912 orientation hypotheses (S_i=4).

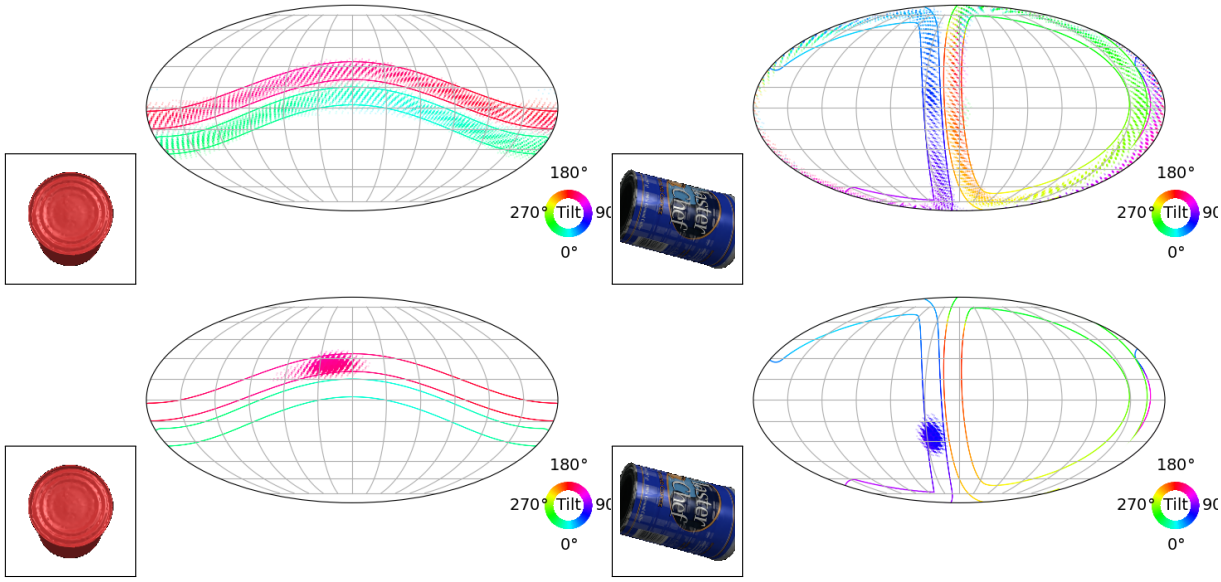

Fig. 9. Comparison of training with pseudo ground-truth generation and single ground-truth orientation on the Uniform (left) and Texture (right) dataset. Top: Learning with pseudo ground-truth orientation. Bottom: Learning with a single ground-truth orientation. The model trained using pseudo ground-truth annotations learns to complete orientation distribution, whereas the model trained using a single ground-truth learns only the single ground-truth but does not learn the complete orientation distribution.

orientation estimate introduced by occlusion. Nevertheless, the model does not learn the correct orientation distribution from a single ground-truth pose.

B. Comparison With Training Using Different Ground-Truths

To evaluate the effectiveness of the pseudo ground-truth pose labeling scheme, we trained the implicit orientation estimation model using three different types of ground-truth poses: single ground-truth pose used to render the image, complete set of *proper symmetry* ground-truth poses generated analytically, and the pseudo ground-truth poses generated using our pipeline. The qualitative results are presented in Fig. 9. In Table I, we report the quantitative comparison results. The single ground-truth model achieves a higher LLH score and a similar MAAD score for all objects, compared to both analytical and pseudo ground-truth models, i.e. the single ground-truth model learns to estimate one single orientation precisely but fails to learn the symmetry orientations, whereas the other two models manage to learn the complete set of symmetry orientations. Moreover, the pseudo ground-truth

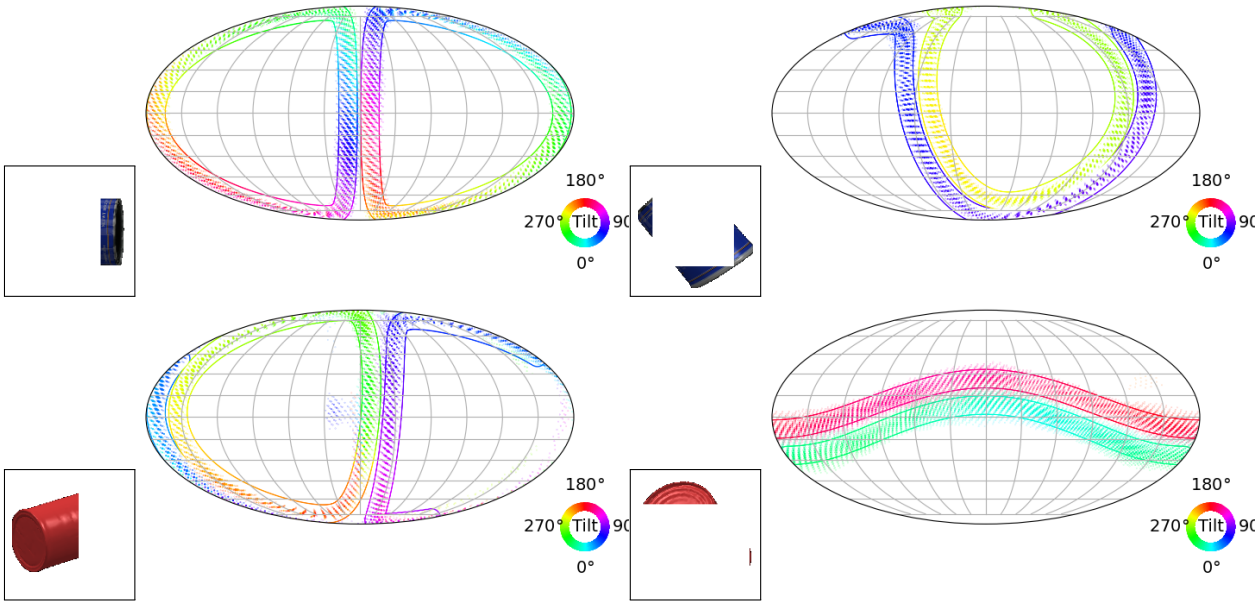

Fig. 10. Orientation distributions predicted by our model in the presence of occlusion. Top: Texture dataset. Bottom: Uniform dataset.

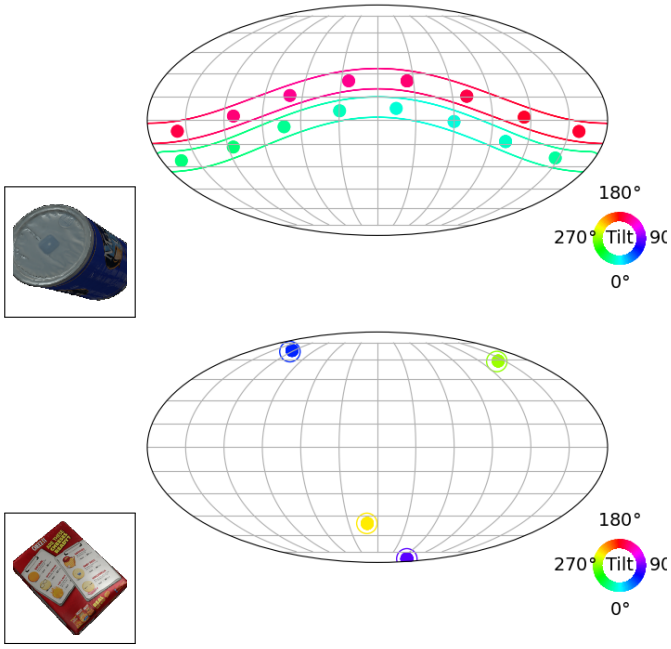

Fig. 11. Visualizing the generated pseudo ground-truth orientation labels. Dots represent the generated pseudo ground-truth orientation labels, whereas the circles and the continuous lines represent the ground-truth orientation. The generated pseudo ground-truth orientation distribution correspond to the ground-truth orientation distribution with high accuracy.

model achieves results similar to the analytical model on all three metrics. Based on these results, we can conclude that the automatic pose labeling scheme is able generate pose labels with high accuracy and covers a sufficiently big portion of the set of *proper symmetries* for the model to learn the symmetries. Furthermore, as a measure of accuracy of the generated pseudo ground-truth orientation labels, we report

TABLE II
EVALUATION OF THE PSEUDO GROUND-TRUTH ORIENTATION LABELS

Object	Dataset	MAAD[°]
can	*Texture*	1.44
	Uniform	3.12
box	*Texture*	4.14
	Uniform	4.3
bowl	*Texture*	1.42
	Uniform	1.89
Average		2.72

in Table II the MAAD metrics of the generated pseudo ground-truth orientation labels for all three objects. Overall, the average error rate of the generated pseudo ground-truth labels is ∼2.7°. Among the three objects present in the dataset, the *box* object has the highest MAAD error. This can be attributed to the fact that *box* exhibits discrete symmetry. Generating pose labels for discrete symmetry is more difficult than for continuous symmetry. As shown in Fig. 11, the pseudo ground-truth pose labels correspond to the ground-truth orientation distribution with a high degree of accuracy.

C. Backbone Ablations

CNN models learn features that generalize well across datasets. However, the degree of generalization varies across different architectures. In order to find the architecture best suited for usage as backbone feature extractor in the ImplicitPDF model, we experimented with ResNet [29], and ConvNeXt [30] architectures. Convolutional neural networks, in general, learn low-level image features at a high resolution in the initial layers and high-level features at low resolution in the final layers [31]–[33]. For many computer vision tasks like object classification and object detection, high-level features—despite being low resolution—are ideal, whereas tasks like

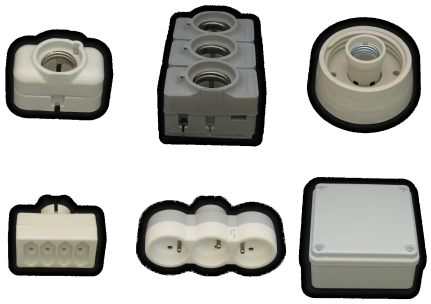

Fig. 12. T-Less dataset objects used for evaluating our method.

semantic segmentation benefit from access to low-level features [34]–[36]. To evaluate the effectiveness of different ResNet layers as feature extractors, we experimented with two different ResNet backbone models—both derived from ResNet-18 and taking images of size 224×224 as input. The first model is ResNet-18 with only the last fully connected layer removed. The features extracted from this model form a vector of size 512. We call this model ResNet-18-Full. The second model is ResNet-18 with the fully connected layer and the last convolutional block removed. The features extracted from this model are of shape 1024×14×14. We call this model ResNet-18-2. Additionally, we also experimented with the ConvNeXt Tiny model [30]. The features extracted from this model form a vector of size 768. In Table III, we present the quantitative results of the comparison. In terms of both LLH and MAAD metrics on the Uniform and Texture datasets, all models perform similarly. The size of the features extracted from the ResNet-Full backbone model is the smallest, though. Thus, we use the ResNet-Full model as the backbone in the remaining experiments.

TABLE III
COMPARISON OF DIFFERENT MODELS AS FEATURE EXTRACTOR.

	Metric	ResNet		ConvNeXt
		18-Full	18-2	
can	LLH	2.35	2.31	3.5
	MAAD	2.45	3.16	3.46
box	LLH	4.45	4.23	4.64
	MAAD	7.48	8.73	8.53
bowl	LLH	4.54	3.21	3.58
	MAAD	8.9	4.82	5.49

D. Evaluation on T-Less Dataset

The T-Less Dataset consists of RGB-D images of texture-less objects of varying sizes along with 6D pose annotations. Training data consists of RGB-D images of individual objects placed in isolation with black background. We evaluate our method on a subset of T-Less objects that exhibit geometric symmetries (shown in Fig. 12). We use the variant of the T-Less dataset proposed by Gilitschenski *et al.* [17] in which the training images provided in the original T-Less is split into training, validation, and test sets. In Table IV, we present quantitative results. Our method achieves a MAAD score of ∼3.6° and an LLH score of 5.23. Since we report the metrics only for the objects that exhibit geometric symmetries, an uncertainty always exists in terms of the object orientation. Thus, our model performs worse in terms of the LLH metrics, compared to other methods, but this does not affect the accuracy in terms of the MAAD metrics. Moreover, compared to the model trained using analytically generated ground-truth orientation labels, which achieves a MAAD score of ∼1.7° and an LLH score of 5.7, our method performs only slightly worse. This indicates that analytically generating ground-truth orientation labels for objects with complex geometric symmetry is not trivial. Moreover, the pose labeling scheme generates pseudo pose labels (SE(3)), whereas the analytical ground-truth generation is possible only for orientation labels (SO(3)). Having access to pose labels enables training complete pose estimation models. Thus, the proposed automated pose labeling scheme serves as an efficient alternative to generating ground-truth orientation labels analytically.

TABLE IV
COMPARISON RESULTS ON T-LESS DATASET.

Method	LLH ↑	MAAD[°] ↓
Deng *et al.* [16]	5.3	23.1
Gilitschenski *et al.* [17]	6.9	3.4
Prokudin *et al.* [37]	8.8	34.3
Murphy *et al.* [2]	9.8	4.1
Analytical	5.7	1.7
Ours*	5.23	3.6

* Results from only a subset of the T-Less objects (shown in Fig. 12).

V. DISCUSSION & CONCLUSION

In this article, we presented the ImplicitPDF model for learning object orientation for symmetrical objects. We proposed a pseudo ground-truth labeling scheme to generate pose annotations and used it to train the ImplicitPDF model without any manual pose annotations. We quantified the advantages of multiple pseudo ground-truth labels over the single ground-truth pose label for training the ImplicitPDF model and the accuracy of the pose labeling scheme. Moreover, by comparing with the models trained using analytically generated ground-truth orientation, we demonstrated the effectiveness of the automatic pose labeling scheme. Overall, our method predicts the complete set of *proper symmetries* for uniform color objects as well as for objects with texture with a high degree of accuracy. In the future, we plan to extend the implicit orientation model ($\mathbf{SO}(3)$) to 6D object pose ($\mathbf{SE}(3)$) to include the estimation of the translational component and reason about its uncertainty.

ACKNOWLEDGMENT

This work has been funded by the German Ministry of Education and Research (BMBF), grant no. 01IS21080, project Learn2Grasp: Learning Human-like Interactive Grasping based on Visual and Haptic Feedback.

References

[1] R. Brégier, F. Devernay, L. Leyrit, and J. L. Crowley, "Defining the pose of any 3D rigid object and an associated distance," *International Journal of Computer Vision (IJCV)*, vol. 126, pp. 571–596, 2018.

[2] K. A. Murphy, C. Esteves, V. Jampani, S. Ramalingam, and A. Makadia, "Implicit-PDF: Non-parametric representation of probability distributions on the rotation manifold," in *International Conference on Machine Learning (ICML)*, PMLR, 2021, pp. 7882–7893.

[3] Y. Xiang, T. Schmidt, V. Narayanan, and D. Fox, "PoseCNN: A convolutional neural network for 6D object pose estimation in cluttered scenes," in *Robotics: Science and Systems (RSS)*, 2018.

[4] T. Hodaň, M. Sundermeyer, B. Drost, Y. Labbé, E. Brachmann, F. Michel, C. Rother, and J. Matas, "BOP challenge 2020 on 6D object localization," in *European Conference on Computer Vision (ECCV)*, 2020, pp. 577–594.

[5] Y. Labbe, J. Carpentier, M. Aubry, and J. Sivic, "CosyPose: Consistent multi-view multi-object 6D pose estimation," in *European Conference on Computer Vision (ECCV)*, 2020.

[6] Y. Xu, K.-Y. Lin, G. Zhang, X. Wang, and H. Li, "RNNPose: Recurrent 6-DoF object pose refinement with robust correspondence field estimation and pose optimization," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 14 880–14 890.

[7] A. Amini, A. S. Periyasamy, and S. Behnke, "YOLOPose: Transformer-based multi-object 6D pose estimation using keypoint regression," in *International Conference on Intelligent Autonomous Systems (IAS)*, 2022.

[8] G. Pitteri, M. Ramamonjisoa, S. Ilic, and V. Lepetit, "On object symmetries and 6D pose estimation from images," in *International Conference on 3D Vision (3DV)*, IEEE, 2019.

[9] A. S. Periyasamy, M. Schwarz, and S. Behnke, "Robust 6D object pose estimation in cluttered scenes using semantic segmentation and pose regression networks," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2018.

[10] C. Esteves, A. Sud, Z. Luo, K. Daniilidis, and A. Makadia, "Cross-domain 3D equivariant image embeddings," in *International Conference on Machine Learning (ICML)*, PMLR, 2019, pp. 1812–1822.

[11] A. Saxena, J. Driemeyer, and A. Y. Ng, "Learning 3D object orientation from images," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2009, pp. 794–800.

[12] M. Rad and V. Lepetit, "BB8: A scalable, accurate, robust to partial occlusion method for predicting the 3D poses of challenging objects without using depth," in *IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 3828–3836.

[13] E. Corona, K. Kundu, and S. Fidler, "Pose estimation for objects with rotational symmetry," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2018.

[14] M. Sundermeyer, Z.-C. Marton, M. Durner, and R. Triebel, "Augmented autoencoders: Implicit 3D orientation learning for 6D object detection," *International Journal of Computer Vision (IJCV)*, vol. 128, no. 3, pp. 714–729, 2020.

[15] 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 *IEEE/CVF International Conference on Computer Vision (CVPR)*, 2019.

[16] H. Deng, M. Bui, N. Navab, L. Guibas, S. Ilic, and T. Birdal, "Deep bingham networks: Dealing with uncertainty and ambiguity in pose estimation," *International Journal of Computer Vision (IJCV)*, pp. 1–28, 2022.

[17] I. Gilitschenski, R. Sahoo, W. Schwarting, A. Amini, S. Karaman, and D. Rus, "Deep orientation uncertainty learning based on Bingham loss," in *International Conference on Learning Representations (ICLR)*, 2019.

[18] D. Mohlin, J. Sullivan, and G. Bianchi, "Probabilistic orientation estimation with matrix Fisher distributions," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, 2020, pp. 4884–4893.

[19] B. Okorn, M. Xu, M. Hebert, and D. Held, "Learning orientation distributions for object pose estimation," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2020.

[20] Y. Xie, T. Takikawa, S. Saito, O. Litany, S. Yan, N. Khan, F. Tombari, J. Tompkin, V. Sitzmann, and S. Sridhar, "Neural fields in visual computing and beyond," in *Computer Graphics Forum*, Wiley Online Library, vol. 41, 2022, pp. 641–676.

[21] L. Mescheder, M. Oechsle, M. Niemeyer, S. Nowozin, and A. Geiger, "Occupancy networks: Learning 3D reconstruction in function space," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[22] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "Nerf: Representing scenes as neural radiance fields for view synthesis," *Communications of the ACM*, vol. 65, no. 1, pp. 99–106, 2021.

[23] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, 2017.

[24] K. M. Gorski, E. Hivon, A. J. Banday, B. D. Wandelt, F. K. Hansen, M. Reinecke, and M. Bartelmann, "HEALPix: A framework for high-resolution discretization and fast analysis of data distributed on the sphere," *The Astrophysical Journal*, vol. 622, no. 2, p. 759, 2005.

[25] D. W. Lyons, "An elementary introduction to the Hopf Fibration," *Mathematics Magazine*, vol. 76, no. 2, pp. 87–98, 2003.

[26] Q.-Y. Zhou, J. Park, and V. Koltun, "Fast global registration," in *European Conference on Computer Vision (ECCV)*, Springer, 2016, pp. 766–782.

[27] R. B. Rusu, N. Blodow, and M. Beetz, "Fast point feature histograms (FPFH) for 3D registration," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2009, pp. 3212–3217.

[28] V. Makoviychuk, L. Wawrzyniak, Y. Guo, M. Lu, K. Storey, M. Macklin, D. Hoeller, N. Rudin, A. Allshire, A. Handa, and G. State, "Isaac Gym: High performance GPU based physics simulation for robot learning," in *Thirty-fifth Conference on Neural Information Processing Systems (NeurIPS)*, 2021.

[29] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE/CVF International Conference on Computer Vision (CVPR)*, 2016.

[30] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A ConvNet for the 2020s," in *IEEE/CVF International Conference on Computer Vision (CVPR)*, 2022.

[31] S. Behnke, *Hierarchical neural networks for image interpretation*. LNCS, Springer, 2003, vol. 2766.

[32] H. Schulz and S. Behnke, "Deep learning - Layer-wise learning of feature hierarchies," *Künstliche Intelligenz (Artificial Intelligence)*, vol. 26, no. 4, pp. 357–363, 2012.

[33] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[34] G. Lin, A. Milan, C. Shen, and I. Reid, "RefineNet: Multi-path refinement networks for high-resolution semantic segmentation," in *IEEE/CVF International Conference on Computer Vision (CVPR)*, 2017, pp. 1925–1934.

[35] M. Schwarz, A. Milan, A. S. Periyasamy, and S. Behnke, "Rgb-d object detection and semantic segmentation for autonomous manipulation in clutter," *International Journal of Robotics Research*, vol. 37, no. 4-5, pp. 437–451, 2018.

[36] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, Springer, 2015, pp. 234–241.

[37] S. Prokudin, P. Gehler, and S. Nowozin, "Deep directional statistics: Pose estimation with uncertainty quantification," in *European Conference on Computer Vision (ECCV)*, 2018, pp. 534–551.