Uncertainty Quantification with Deep Ensembles for 6D Object Pose Estimation

Kira Wursthorn [*], Markus Hillemann, Markus Ulrich

Institute of Photogrammetry and Remote Sensing (IPF), Karlsruhe Institute of Technology (KIT), Germany
{kira.wursthorn, markus.hillemann, markus.ulrich}@kit.edu

KEY WORDS: Deep Learning, Uncertainty Quantification, Deep Ensembles, 6D Object Pose Estimation, Reliability

ABSTRACT:

The estimation of 6D object poses is a fundamental task in many computer vision applications. Particularly, in high risk scenarios such as human-robot interaction, industrial inspection, and automation, reliable pose estimates are crucial. In the last years, increasingly accurate and robust deep-learning-based approaches for 6D object pose estimation have been proposed. Many top-performing methods are not end-to-end trainable but consist of multiple stages. In the context of deep uncertainty quantification, deep ensembles are considered as state of the art since they have been proven to produce well-calibrated and robust uncertainty estimates. However, deep ensembles can only be applied to methods that can be trained end-to-end. In this work, we propose a method to quantify the uncertainty of multi-stage 6D object pose estimation approaches with deep ensembles. For the implementation, we choose SurfEmb as representative, since it is one of the top-performing 6D object pose estimation approaches in the BOP Challenge 2022. We apply established metrics and concepts for deep uncertainty quantification to evaluate the results. Furthermore, we propose a novel uncertainty calibration score for regression tasks to quantify the quality of the estimated uncertainty.

1. INTRODUCTION

Determining the 6D pose of an object in its 3D environment, i.e., its 3D orientation and 3D position, from a camera image or from depth sensor data is a fundamental task in computer vision. Applications like augmented reality (Su et al., 2019), vision-assisted robot manipulation (Steger et al., 2018; Ulrich and Hillemann, 2024), bin picking Drost et al. (2017), and autonomous systems as self-driving cars (Yurtsever et al., 2020) rely on accurate object poses estimated from RGB(-D) images. In the real world, complex scenes arise that lead to typical challenges for 6D object pose estimation: The scene might be cluttered with multiple instances of varying object categories, sometimes with more than one instance of a given object category and occluded instances. Objects might be symmetric or contain inter-object similarity, where one object is built from parts of other objects. In addition, object surfaces can be challenging if they are textureless or reflective, for example. The Benchmark for 6D Object Pose Estimation (BOP) and the associated BOP Challenge 2020 (Hodaň et al., 2020) and 2022 (Sundermeyer et al., 2023) cover these challenges and allow a robust evaluation of state-of-the-art methods. These methods incorporate deep learning at large, taking advantage of the ability of deep neural networks to learn complex patterns on sufficient amounts of data and use RGB images as well as depth information to compute the 6D object pose. Many of the top-performing approaches (Park et al., 2019; Labbé et al., 2020; Haugaard and Buch, 2022; Wang et al., 2021) are composed of three major stages: First, an off-the-shelf object detector locates the target object instance in the image of the scene. Second, a deep neural network predicts the 2D–3D correspondences, and third, a variant of the Perspective-n-Point (PnP) algorithm, often combined with RANSAC, provides the 6D pose of the detected object instances. Optionally, depth information is used for pose refinement.

In safety-critical applications like autonomous driving (McAllister et al., 2017) and demanding industrial applications like quality inspection and automation (Heizmann et al., 2022) , the prediction of an object pose is often not sufficient to make informed decisions. Instead, the associated object pose uncertainty must also be taken into account. For example, consider the task of a robot grasping a cup whose pose is estimated based on a RGB(-D) input image that does not show the handle of the cup. This leads to an ambiguous pose estimate. If the robot grasps the cup based on that pose estimate, the object or the robot might be damaged. In combination with a measure of pose uncertainty, this scenario can be identified and prevented by choosing a different camera angle or, in a bin picking application, choosing another object to grasp that has a lower uncertainty. In deep learning, popular uncertainty quantification (UQ) methods include softmax probability (Hendrycks and Gimpel, 2018), Monte-Carlo Dropout (Gal and Ghahramani, 2016), and Deep Ensembles (Lakshminarayanan et al., 2017). While softmax predictions are only used for classification and segmentation tasks, Monte-Carlo Dropout and Deep Ensembles can be applied to regression tasks as well, and therefore are suited for object pose estimation. Deep ensembles of random initialized networks perform best and are more robust under datashift, compared to dropout methods, post-hoc calibration by temperature scaling, and methods motivated by Bayesian inference (Ovadia et al., 2019).

The application of UQ methods to multi-stage approaches for 6D object pose estimation is not straightforward. These methods are usually designed for segmentation and classification tasks, which often are single-stage approaches in the sense that they are end-to-end trainable. Since 6D object poses have one orientation component in $\mathrm{SO}(3)$ and one position component in $\mathbb{R}^3$, the object pose is often considered in a decoupled fashion, handling orientation and position separately. While it can be generally assumed that the object position in $\mathbb{R}^3$ is normally distributed, modelling the orientation distribution is more complex. Considering this, Deep Ensembles and UQ methods that draw samples from the posterior predictive distribution have the

[*] Corresponding author

https://doi.org/10.5194/isprs-annals-X-2-2024-223-2024 | © Author(s) 2024. CC BY 4.0 License.

advantage that no assumptions concerning the underlying distributions have to be made. Surprisingly, up to now, there is no work that uses a deep ensemble approach for UQ in object pose estimation. The most closely related method by Shi et al. (2021) uses two to three heterogeneous pretrained pose estimation models to estimate the pose disagreement. However, while this approach reduces the computational cost of training and inferring a large ensemble of models, this approach diverges from the deep ensemble methodology and does not produce uncertainty estimates.

In this work, we propose a method to quantify the uncertainty of multi-stage 6D object pose estimation approaches with the current state-of-the-art deep learning UQ method, namely deep ensembles. For the implementation, we choose SurfEmb (Haugaard and Buch, 2022), a top-performing 6D object pose estimation method. We evaluate the estimated pose results and their uncertainties using reliability diagrams and BOP metrics on the T-LESS (Hodaň et al., 2017) and YCB-V (Xiang et al., 2018) benchmark datasets for object pose estimation. Furthermore, we introduce a novel metric for the evaluation of uncertainty estimates in regression tasks in general.

2. RELATED WORK

As UQ in deep learning and explainable AI gain more and more interest, works on the reliability of both network predictions and estimated uncertainties have increased in the recent years. In Section 2.1, an overview of popular UQ approaches in deep learning in general is given. In Section 2.2, works on object pose uncertainties and pose distributions are described.

2.1 UQ in Deep Learning

UQ in deep learning often distinguishes between different types of uncertainties depending on their source. The predictive uncertainty is often split into aleatoric and epistemic uncertainty. Aleatoric uncertainty captures the uncertainty that is inherent in the input data, e.g. noise in an image, while epistemic uncertainty refers to a lack of knowledge, i.e. the uncertainty of the network parameters (Kendall and Gal, 2017).

Deep-learning-based approaches for object pose estimation integrate large deep neural networks in their pipelines in most cases. Consequently, these networks consist of a large amount of parameters and non-linearities that make the computation of the exact posterior probability distributions of the network’s predicted outputs generally intractable (Blundell et al., 2015; Loquercio et al., 2020). As a consequence, approximation approaches are used for UQ. Approaches based on Bayesian inference transform common deterministic networks into stochastic ones by placing probability distributions over either the activations and/or the weight parameters (Jospin et al., 2022), leading to Bayesian neural networks (BNNs). While BNNs have a mathematically sound foundation, the high parameter counts of deep neural networks make a direct solution impossible.

Bayes by Backprop (Blundell et al., 2015) is one work proposing variational inference to learn the parameters of approximate distributions over the weights. At inference time, weights are sampled from the learned distributions resulting in an ensemble of networks that is used to sample the posterior distribution of the predictions. Because BNNs come with a high computational cost, Monte-Carlo Dropout (Gal and Ghahramani, 2016) was proposed where dropout regularization (Srivastava et al., 2014) at inference time approximates a stochastic Gaussian process. Then, the posterior predictive distribution is sampled from multiple forward passes through networks with varying dropout masks requiring additional runtime at inference time. Furthermore, Monte-Carlo Dropout capture only the epistemic uncertainty and, therefore, has been combined with probabilistic networks (Kendall and Gal, 2017) and assumed density filtering (Gast and Roth, 2018; Loquercio et al., 2020) for predictive UQ. Another drawback of Monte-Carlo Dropout is that the estimated uncertainties need to be calibrated (Gal and Ghahramani, 2016). In contrast, Deep Ensembles do not require a calibration (Lakshminarayanan et al., 2017). They present a robust way to estimate predictive uncertainty in computer vision tasks such as classification, semantic segmentation, and depth estimation, and are considered state-of-the-art in deep learning UQ (Ovadia et al., 2019; Gustafsson et al., 2020; Wursthorn et al., 2022). Ensemble distillation can be used to overcome the high memory costs of ensembles while achieving comparable uncertainty results (Landgraf et al., 2023). Recently, (Mukhoti et al., 2023) proposed a deterministic UQ approach that provides similar results to deep ensembles, even on out-of-distribution examples.

2.2 UQ for Object Pose Estimation

Despite the importance of reliable pose estimates, there are few works that focus on UQ in the context of 6D object pose estimation (Thalhammer et al., 2023a). Often, UQ for object pose estimation is referred to as the estimation of a pose distribution. Many works use Bingham distributions to model the orientation distribution (Gilitschenski et al., 2020; Okorn et al., 2020; Deng et al., 2022; Sato et al., 2022). Gilitschenski et al. (2020) present a new Bingham loss function for orientation distribution learning and Okorn et al. (2020) propose two methods to quantify the uncertainty of orientations for non-symmetric and symmetric objects, respectively. The first method uses an isotropic Bingham distribution to model orientation distribution while the latter learns a multi-modal non-parametric distribution. Deng et al. (2022) propose Deep Bingham Networks as UQ framework by considering a family of pose hypotheses. Sato et al. (2022) present a simple way how a prediction head estimating the parameters of a Bingham distribution can be incorporated into PoseCNN (Xiang et al., 2018). Manhardt et al. (2019) use the distribution of pose hypotheses to handle object ambiguities, a goal that was also of interest in Deng et al. (2021) where the orientation distribution is considered while tracking object poses in video frames. In turn, Jeon et al. (2023) use the object ambiguities to estimate confidences for keypoint selection. Further approaches for object pose estimation leverage keypoint confidences to improve the performance and to provide a measure of reliability of the pose estimates (Peng et al., 2019; Huang et al., 2022; Yang and Pavone, 2023). Others estimate confidences for the pose hypotheses to increase the accuracy of the final average pose result (Hu et al., 2019; Thalhammer et al., 2023b). Recent works use non-parametric distributions to implicitly model the pose distribution in $\mathrm{SE}(3)$. Haugaard et al. (2023) present an efficient way to learn pose distributions at different resolution levels. Recently, Zhou et al. (2023) combine SurfEmb with inverse graphics and provide a log-likelihood scoring for the estimated poses. In context of the UQ methods mentioned in Section 2.1, these approaches for object pose distribution estimation can be considered as single deterministic approaches to pose UQ. In contrast to sampling-based UQ methods, single deterministic approaches do not require multiple forward passes at inference time. However, they are sensitive to the underlying network architecture, training procedure, and training data (Gawlikowski et al., 2023). Ensemble methods like deep ensembles have been shown to be more robust under datashift and outperform other methods like Monte-Carlo Dropout (Ovadia et al., 2019). Furthermore, not all mentioned object pose estimation approaches leveraging uncertainties in their methodology offer uncertainties of the final pose results Brachmann et al. (2016). Also, the quality of the uncertainty estimates is often not explicitly evaluated. In addition, the incorporation of uncertainties mostly comes with complex changes in established pose estimation methodologies. In contrast, deep ensembles offer a simple approach to UQ. In this paper, we show how a deep-learning-based object pose estimation approach is extended to additionally quantify uncertainties with deep ensembles.

3. BACKGROUND OF SURFEMB AND DEEP ENSEMBLES

Given a multi-stage 6D object pose estimation approach, we evaluate the applicability of the UQ method of deep ensembles to the task of deep 6D object pose estimation. First, in Section 3.1, an introduction to the components of the SurfEmb (Haugaard and Buch, 2022) method is given, which is chosen as the exemplary method for the task of 6D object pose estimation. Section 3.2 explains the prerequisites that have to be fulfilled by the ensemble baseline models in order to ensure the creation of a well-calibrated deep ensemble.

3.1 SurfEmb

We conduct our experiments using SurfEmb, which is in the top ten of the best performing methods in the BOP challenge 2022 (Sundermeyer et al., 2023). Like many 6D object pose estimation methods, SurfEmb is a multi-stage approach, which is why the insights gained in this work can be transferred to similar multi-stage approaches as well. It uses the 2D object instance detections produced by CosyPose (Labbé et al., 2020) with Mask R-CNN (He et al., 2017) and trains a deep neural network to predict 2D–3D correspondences that are forwarded to a PnP algorithm that estimates the object poses. More specifically, SurfEmb learns dense and continuous 2D–3D correspondence distributions by using high-dimensional embeddings of the object surface coordinates. The correspondence network is trained in a self-supervised fashion using a contrastive loss. The positive and negative training examples are provided by the so-called key model, a sinusoidal representation network (SIREN) MLP (Sitzmann et al., 2020) that transforms a 3D object surface coordinate into a 12D embedding space. The correspondence network or the so-called query model with a U-Net (Ronneberger et al., 2015) architecture and a ResNet18 (He et al., 2016) backbone is then trained to predict pixel-wise 12D surface embeddings from an input image crop of an object instance. The 2D–3D correspondences are then used in AP3P (Ke and Roumeliotis, 2017), an algebraic P3P algorithm, to obtain object pose hypotheses followed by a pose hypotheses scoring. Pose hypotheses that exceed a score threshold are locally refined and can be further refined with depth data.

3.2 Deep Ensembles

Lakshminarayanan et al. (2017) offer a recipe with three prerequisites that baseline models must fullfill in order to obtain a well-calibrated deep ensemble of models that predicts accurate uncertainties. The prerequisites concern i) the model weights initialization scheme, ii) the scoring rule used during model training, and iii) whether a form of adversarial training is applied.

Model Weights Initialization Scheme. The weights of each model in the ensemble are initialized randomly. The randomness causes the models to reach different modes of the loss during training and thus the ensemble can better cover the posterior distribution of predictions. This is one of the main reasons why deep ensembles work well in practice (Fort et al., 2020).

Scoring Rule. For network training, a probabilistic scoring rule, i.e. a scoring rule that quantifies the quality of the predicted probability distributions, is required. In general, for classifiers that use loss functions maximizing a likelihood, like softmax cross entropy, this is already satisfied. In regression tasks, a Gaussian negative log-likelihood (NLL) can be used.

Adversarial Training. Deep Ensembles incorporate adversarial training (Goodfellow et al., 2015) for predictive distribution smoothing. During each training step, negative examples are generated based on the current input, which are included in the calculation of the loss. Adversarial Training is proposed as an optional step during model training (Lakshminarayanan et al., 2017).

Each individual model in the ensemble must fulfill the first two prerequisites. Let N be the ensemble size, i.e. the number of ensemble members that are trained in accordance with the above prerequisites and used to produce the ensemble results at inference time. In Ovadia et al. (2019), it was shown that an ensemble size of $N = 5$ produces good results. Nevertheless, N is an empirical value and should be determined for each application specifically. The higher the ensemble size, the better the underlying posterior predictive distribution of the ensemble outputs can be approximated.

4. METHODOLOGY

In this section, we describe how we apply deep ensembles to SurfEmb and how we evaluate it. In Section 4.1, the prerequisites described in Section 3.2 are checked and applied to SurfEmb. The evaluation methodology is described in Section 4.2.

4.1 SurfEmb Deep Ensemble

In the following, we explain how the three prerequisites described in Section 3.2 are taken into account when using an ensemble of SurfEmb. Next to these prerequisites, the ensemble size N must be defined empirically for our application, which is discussed in the experiments in Section 5.

Model Weights Initialization Scheme. In the original publication of SurfEmb, the ResNet18 backbone of the U-Net architecture of the query model is initialized with pre-trained weights on ImageNet (Deng et al., 2009). Instead, according to the deep ensemble recipe of Section 3.2, we randomly initialize each ensemble query model with different weights drawn from a normal distribution scheme according to He et al. (2015). As mentioned in Section 3.1, the key model that provides the targets during training is based on a SIREN MLP. Due to the sensitivity of the used sine non-linearities, the SIREN MLP requires the specification of lower and upper bounds of a uniform distribution from which the weights are randomly drawn. By randomly initializing the key models in the ensemble, the models learn different realizations of the 12D embedding space of correspondence distributions.

Scoring Rule. The query model of SurfEmb is trained with a combined loss for the predicted visible object mask and the surface embedding. The object mask is scored by the binary cross entropy while multi-class cross entropy is used as a scoring rule for the surface embeddings. As cross entropy can be considered a proper scoring rule, no modifications are required.

Adversarial Training. SurfEmb does not incorporate an adversarial training. However, it is trained in a self-supervised manor with a contrastive loss taking both negative and positive examples into account. This training regime has a similar effect on the ensemble results as adversarial training. Therefore, and because this prerequisite is optional, we do not perform any further predictive distribution smoothing.

The resulting SurfEmb ensemble consists of $N = 10$ independent query models of whom each model generates a query, and, based on that, produces an object pose estimate at inference time, forming the ensemble pose estimates. One major advantage of an ensemble approach for UQ is that no assumptions are made about the underlying distribution of predictions. In case of object pose estimation, this especially presents an advantage over UQ methods that predict the parameters of distributions explicitly. As a drawback, one has to overcome the challenging endeavor of extracting meaningful, application specific pose uncertainties. In case of object symmetries, it is not guaranteed that all pose estimates of the ensemble members refer to the same object symmetry axis. For this purpose, we select the pose in the ensemble prediction with the highest score as reference and align all $N - 1$ other poses to that reference based on the known symmetric transformations of the object model.

4.2 Ensemble Evaluation

The pose estimator ensemble is evaluated on the test set of the corresponding training dataset. Let the test dataset D be defined as $D = \{(x_t, y_t)\}_{t=1}^{T}$ where x_t is the t-th input data point, i.e. the image crop of an object instance, and y_t the corresponding annotated ground truth pose of the depicted object instance in the camera coordinate frame, composed of a 3D rotation matrix $\mathbf{R}$ and a translation vector $\mathbf{t}$. For the t-th entry in the test dataset, the n-th ensemble member H_n, with $n = \{1, 2, ..., N\}$, outputs a prediction, resulting in a sample of N predictions that are drawn from the posterior predictive distribution. An approximate distribution can be fit to this predictive distribution to get ensemble results that consist of the parameters of the approximate distribution. In the simplest case, a Gaussian distribution is defined by the mean μ_t and the standard deviation σ_t of the ensemble predictions on the t-th dataset entry. Despite the assumption made about the underlying posterior predictive distribution, the standard deviation offers the advantage of being easy to interpret.

4.3 Uncertainty Evaluation

For a consistent evaluation of the ensemble's means and standard deviations, we compute reliability diagrams or calibration plots (Lakshminarayanan et al., 2017; Kuleshov et al., 2018). The reliability diagram for an ensemble or any forecaster H that predicts a cumulative density function (CDF) F_t of the t-th dataset entry is computed based on the following assumption: H is well calibrated on dataset D if the predicted CDFs match the empirical CDFs when the dataset size T tends towards infinity (Kuleshov et al., 2018). Given M chosen expected confidence levels $0 \leq p_1 < p_2 < ... < p_M \leq 1$, the corresponding observed confidence level $\hat{p}_j$ to each threshold p_j is calculated by computing the empirical frequency (Kuleshov et al., 2018):

$$\hat{p}_j = \frac{|\{y_t | F_t(y_t) \leq p_j, t = 1, ..., T\}|}{T} \quad . \tag{1}$$

If H is well calibrated, the values $\{(p_j, \hat{p}_j)\}_{j=1}^{M}$ form a straight line which passes through the origin and has slope 1 (Kuleshov et al., 2018). In our case, where we estimate the probability density function in terms of the mean and standard deviation of the ensemble predictions on the t-th dataset entry, we first compute the corresponding predicted CDF F_t and get the empirical CDF by drawing the t-th target value y_t from F_t.

Based on this reliability diagram, we propose an uncertainty evaluation metric that takes into account the area between the perfect calibration where the target and predictive distributions match and the actually observed confidence levels. We call the metric the uncertainty calibration score (UCS). Whereas in case of a perfect uncertainty calibration the area is zero, the worst case uncertainty calibration corresponds to the possible maximum value of this area which is $A_{max} = 0.25$. Therefore, based on these lower and upper bounds, we propose to quantify the calibration quality by

$$\mathrm{UCS} = 1 - A/A_{max} \quad , \tag{2}$$

where A is the area estimated from the reliability diagram and calculated by using the composite trapezoidal rule:

$$A = \int_0^1 f(p)\, dp \approx \frac{\Delta p}{2} \sum_{j=1}^{M} \left(f(p_{j-1}) + f(p_j)\right) \quad , \tag{3}$$

where $f(p) = |\hat{p} - p|$ is the absolute difference between the observed confidence level $\hat{p}_j$ estimated using Equation (1) and the expected confidence level p_j. The computation increment is set to $\Delta p = 0.1$. UCS is bound between $[0, 1]$, where a higher value indicates a better calibration. Consequently, the metric is easy to interpret and facilitates a comparison or even a ranking of the different methods. The estimated area A can be interpreted as the calibration error and is similar to the expected calibration error (ECE) (Guo et al., 2017), a popular network calibration error metric for semantic segmentation and classification tasks. Both, the ECE and A in Equation (2) are computed based on the differences between the observed accuracy or confidence and the predicted or expected confidence. In the sense that the ECE depends on the number of bins, UCS also depends on the chosen computation increment Δp in the calculation of the observed confidence levels of the reliability diagram in Equation (1). On the other hand, this dependency implies flexibility with respect to the choice of the assumed underlying distribution. In combination with sampling-based UQ methods, UCS can be used to find the optimal sample size. Note that in contrast to the ECE, UCS can be applied to any calibration plot, regardless whether the underlying task is regression, classification, or segmentation. The parameterization of the cumulative distribution, which is assumed for the calculation of the observed confidence levels, can be freely chosen, since in the case of deep ensembles the choice is not restricted. Therefore, varying distributions for the orientation and position component can be considered in the computation of UCS for a deep ensemble of an object pose estimator. Here, we use Gaussians to parameterize the distribu-

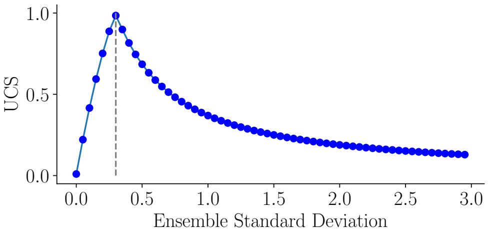


Figure 1. UCS for simulated uncertainty predictions with a ground truth standard deviation $\sigma_{true} = 0.3$ (dashed line). For a perfect calibration, where the simulated and predicted uncertainty match, UCS is close to 1.

tions of both the orientation and position.

Since the ground truth posterior uncertainty of the ensemble outputs is unknown, we use synthetic data to evaluate the performance of our proposed quality score UCS. Given a dataset size of $T = 10000$, we sample each t-th target from a uniform distribution. The corresponding output distributions are represented by a mean and the standard deviation σ_{pred}. For each target, the ensemble output mean is sampled from a Gaussian distribution centered on the target and with a standard deviation of $\sigma_{true} = 0.3$. Based on the reliability diagram, we expect that UCS $= 100.0\,\%$, if $\sigma_{true} = \sigma_{pred}$ for all T dataset entries. The more σ_{pred} differs from the expected value of σ_{true}, the more should UCS decrease. The results of the simulation are shown in Figure 1 and confirm our expectations. In case that the predicted uncertainty matches the ground truth standard deviation, the UCS is 0.990. For wrong predictions the UCS is smaller. As the standard deviations do not have an upper bound, UCS slowly converges to zero if the estimated uncertainties are too large. In case of a few outliers ($0.01\,\%$ of T) that successfully target the ground truth but where $\sigma_{pred} \neq \sigma_{true}$, UCS is not affected significantly and therefore proves to be robust to outliers.

5. EXPERIMENTS

We conducted experiments on the T-LESS (Hodaň et al., 2017) and YCB-V (Xiang et al., 2018) datasets. Both are part of the BOP challenge and include photorealistic rendered training images of randomly sampled cluttered scenes that were added to the datasets as part of the BOP challenge 2020 (Hodaň et al., 2020) and are used to train the models. While T-LESS consists of 30 industrial parts that are largely textureless and in many cases symmetric, the YCB-V dataset contains 21 objects of daily life. Both datasets include CAD object models. The BOP test dataset of T-LESS is composed of 20 cluttered scenes, for each of which 50 real test images are provided. Two example RGB images are shown in Figure 2. The BOP test images of YCB-V are sampled from 12 of the 92 video scenes of the original dataset. Because not all annotated ground truth targets are visible, we only consider target objects where at least $10\,\%$ of the object instance surface is visible and the visible object instance part is represented by at least 1024 pixels in the image. This results in 6423 and 4121 valid ground truth samples for T-LESS and YCB-V, respectively. As these criteria are also used during model training, this test data subset can be considered as in-domain. In contrast, test data points of object instances that are visible by less than $10\,\%$ and whose visible surface masks have fewer than 1024 pixels form an out-of-domain test dataset. In Section 5.1, we evaluate the quality of the pose estimates of the trained ensemble members for T-LESS and YCB-V on their respective BOP test datasets. In Section 5.2, the estimated ensemble distributions are analyzed.

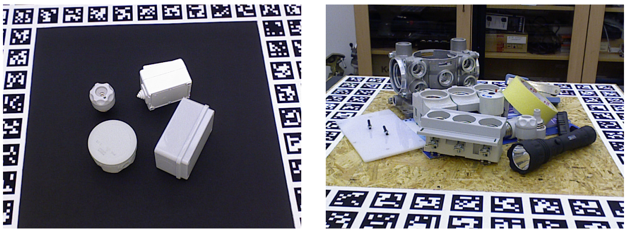

(a) Scene 1 (b) Scene 18

Figure 2. Two examples of RGB images from different scenes of the BOP test dataset of T-LESS.

5.1 Evaluation of the Ensemble Pose Estimates

To ensure that the overall quality of the pose estimates of the trained ensembles is not affected by the random initialization, we evaluated each of the ensemble members, which we call the baseline models, separately. We also evaluated the pose average as the mean over all ensemble members. For evaluation, we applied the BOP error metrics $MSPD$, $MSSD$, and VSD Hodaň et al. (2020). In Table 1, the results with and without depth refinement are compared to the scores of the reproduced SurfEmb model that is provided by the authors, the mean of the scores achieved by the ensemble baseline models, and the scores for the estimated mean poses of the ensemble are reported. Both ensembles trained on T-LESS and YCB-V consist of ten ensemble members each. The scores are defined as the average recall (AR) of the BOP error metrics $MSPD$, $MSSD$, and VSD. While the $MSPD$ error measures the perceivable discrepancy and, therefore, is relevant for augmented reality applications, the $MSSD$ error measures the maximum pose error in the 3D space and is especially relevant for robot manipulation (Hodaň et al., 2020). Both metrics take object symmetries into account. The VSD is the visual surface discrepancy Hodaň et al. (2020). Surprisingly, the poses of the randomly initialized ensemble members achieve similar scores as the ones estimated by the provided SurfEmb models for T-LESS and YCB-V that were initialized with pretrained weights on ImageNet. It seems that in this case, pretraining does not improve the quality of the predictions. Furthermore, it can be observed that ensembling the poses slightly improves the quality of the prediction, a phenomenon that is often taken advantage of in knowledge distillation, where the knowledge of an ensemble is compressed into a single model to overcome the ensemble drawback of high computational costs Hinton et al. (2015).

5.2 Evaluation of the Ensemble Uncertainty

To eliminate possible influences of the 2D object detection on the evaluation of the pose estimation ensemble, the evaluation of the ensemble results is done on object instance image crops based on the ground truth. In Figure 3, the reliability diagram described in Section 4 is shown for the T-LESS ensemble query model outputs, meaning the 12D embedded dense correspondence distributions. The reliability diagram can be interpreted as follows: For instance, with an expected confidence level of 0.10, we observe an actual confidence level of 0.11, meaning that for 11 % of the T-LESS test data points the CDFs of the Gaussian distributions output a probability of $\leq$ 10 %. The Gaussian distributions are parametrized by the ensemble means

		RGB			RGB-D		
Dataset	Scores	Pretrained	Baseline	Ensemble	Pretrained	Baseline	Ensemble
T-LESS	$\mathrm{AR}_{MSPD}\uparrow$	0.846	0.844 ± 0.002	$\mathbf{0.851}$	0.851	0.848 ± 0.002	$\mathbf{0.855}$
	$\mathrm{AR}_{MSSD}\uparrow$	0.554	0.556 ± 0.010	$\mathbf{0.586}$	0.821	0.815 ± 0.003	$\mathbf{0.823}$
	$\mathrm{AR}_{VSD}\uparrow$	0.525	0.527 ± 0.010	$\mathbf{0.561}$	0.784	0.779 ± 0.002	$\mathbf{0.790}$
	$\mathrm{AR}\uparrow$	0.642	0.642 ± 0.007	$\mathbf{0.666}$	0.818	0.814 ± 0.002	$\mathbf{0.823}$
YCB-V	$\mathrm{AR}_{MSPD}\uparrow$	0.759	0.750 ± 0.002	$\mathbf{0.764}$	0.789	0.780 ± 0.003	$\mathbf{0.794}$
	$\mathrm{AR}_{MSSD}\uparrow$	$\mathbf{0.498}$	0.483 ± 0.008	0.489	0.850	0.844 ± 0.003	$\mathbf{0.852}$
	$\mathrm{AR}_{VSD}\uparrow$	$\mathbf{0.433}$	0.417 ± 0.008	0.421	0.757	0.748 ± 0.003	$\mathbf{0.762}$
	$\mathrm{AR}\uparrow$	$\mathbf{0.563}$	0.550 ± 0.005	0.558	0.799	0.791 ± 0.003	$\mathbf{0.803}$

Table 1. Evaluation results of the differently trained SurfEmb models on the BOP test datasets of T-LESS and YCB-V, both without (RGB) and with depth refinement (RGB-D). Shown are the reproduced AR of the models trained and provided by the authors of SurfEmb (Pretrained), the mean AR of the randomly initialized ensemble members (Baseline), and the evaluation results of the mean poses of the ensembles (Ensemble).

and standard deviations and evaluated on the targets. The observed confidences of the ensemble queries are estimated based on the pixel-wise mean and the standard deviation of each embedding dimension. As the targets produced by the key models and the predicted queries are not in the same range, they were normalized with their minimum and maximum values, respectively. The ensemble outputs are predicted on the ground truth object instance crops of the 6423 visible samples of the T-LESS test dataset. As the area between the plotted curve of the observed confidence levels and the diagonal that represents a perfect calibration is very small, the query model seems to be very well calibrated. Accordingly, it has a high corresponding calibration score of $\mathrm{UCS} = 96.0\ \%$. It can be observed that for low expected confidence levels the observed confidence is slightly larger than the expected value. Analogously, for high expected confidence levels, the observed confidence level is slightly lower than expected. Based on the predicted queries by the ensemble members on the ground truth object instance image crops, the pose results of each ensemble member are computed. Figure 4 shows the reliability diagrams of the estimated orientation in form of rotation matrices and position components on T-LESS and YCB-V. It can be observed that, overall, the T-LESS ensemble seems to be better calibrated than the YCB-V ensemble. In Figure 4a shows that the quality of the calibration of the orientation component decreases with the local refinement step, both on T-LESS and YCB-V. The unrefined poses achieve a UCS of $88.5\ \%$ and $80.6\ \%$ while the refined estimates score $79.6\ \%$ and $69.3\ \%$ on T-LESS and YCB-V, respectively. In contrast, the position component is only lightly affected by the local refinement and decreases the UCS of the unrefined estimates by $3.1\ \%$ on T-LESS and by $0.8\ \%$ on YCB-V. While the depth refinement does not influence the orientation estimates, it improves the quality of the estimated position component that also affects its calibration, as it is shown in Figure 4b. While the UCS of the position component on YCB-V increases from $33.6\ \%$ to $44.3\ \%$ when they are refined with the depth data, the UCS on T-LESS decreases from $65.2\ \%$ to $57.6\ \%$. The reason behind this will be part of future work.

In Figures 5a and 5b, the reliability diagrams of different orientation representations of the pose estimates on the ground truth image crops of T-LESS and YCB-V are shown. The poses are unrefined so that the influence of the orientation representations can be better observed and other influences are as much reduced as possible. The four chosen representations are quaternions, Euler angles, Rodriguez' axis-angle representation, and the rotation matrix. These representations were selected based on their importance in orientation and pose estimation tasks and their interpretability. Out of the four representations, Rodriguez' axis-angle representation achieves the highest UCS with $90.5\ \%$ and $77.1\ \%$ on T-LESS and YCB-V, while the quaternion representation scores the lowest with $76.7\ \%$ and $59.3\ \%$. The calibration of the position component is noticably decreased in comparison to the orientation component, regardless of the representation.

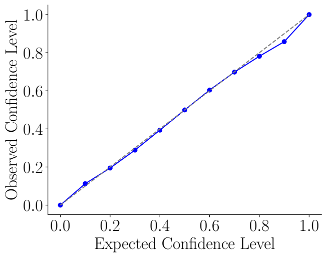
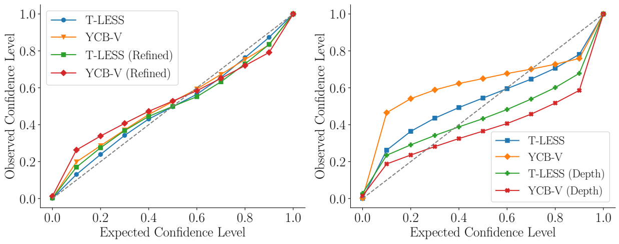


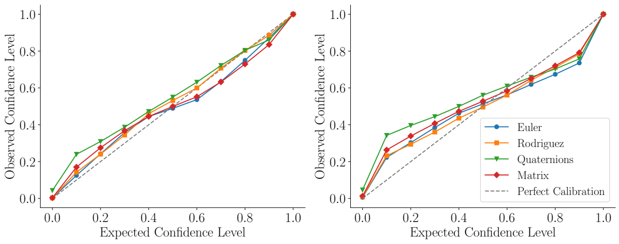
Figure 3. Reliability diagram of the T-LESS query model ensemble with the optimal ensemble size of eight ensemble members. The perfect calibration as the diagonal is represented by the dashed gray line.



(a) Ensemble orientations with and without local refinement

(b) Locally refined ensemble positions with and without depth

Figure 4. Reliability diagrams of the estimated ensemble orientation and position components on T-LESS and YCB-V. The perfect calibration is represented by the dashed gray line.



(a) Different orientation representations on T-LESS

(b) Different orientation representations on YCB-V

Figure 5. Reliability diagrams of the locally refined ensemble orientation components on T-LESS and YCB-V.

6. DISCUSSION

The reliability diagrams of the T-LESS query model ensemble in Figure 3 and of the orientation components on both datasets in Figure 4 show that the deep ensembles are well calibrated. This is also reflected in the high values for UCS. The almost perfect calibration of the query model ensembles is reduced during the follow-up steps of the pose estimation pipeline of SurfEmb. This leads to the conclusion that, while deep ensembles are easy to apply in general, this notion may not be transferred to an ensemble of a pose estimator with multiple stages and a combination of deep learning and algorithms. It may be circumvented by using an end-to-end trainable pose estimator where the PnP algorithm is implemented as part of the trainable network architecture as it was done in GDRNet (Wang et al., 2021) or by using error propagation. However, Figure 3 shows that the ensemble results of the query model on the T-LESS dataset are well calibrated, demonstrating that in this case for UQ with deep ensembles the stage consisting of the 2D object detector does not need to be included. It has to be noted that the position components of the pose estimates is detrimental to the overall calibration. In the reliability diagrams of different orientation representations on T-LESS and YCB-V, shown in Figures 5a and 5b, it can be seen that the choice of the representation has an influence on the imperfections and thus the calibration. Remarkable is the decrease of the calibration quality in case of a representation as quaternions, which may be due to the fact that the assumed normal distribution with one standard deviation value per element is not sufficient.

7. CONCLUSION

In this work, we applied the state-of-the-art deep learning UQ method of deep ensembles to SurfEmb, one of the top-performing multi-stage deep 6D object pose estimation approaches, and evaluated the result on the T-LESS dataset. The adaptation of SurfEmb’s correspondence network to the deep ensemble methodology is straightforward and we find that the ensemble on T-LESS is very well calibrated. However, the following PnP implementation, pose refinement strategies, and pose representations reduce the quality of the estimated predictive uncertainty. Furthermore, we introduced UCS, a novel metric to quantify the estimated uncertainty for regression tasks. UCS is easy to interpret and facilitates a comparison or even a ranking of the different methods. In future work, we want to extend the experiments to other pose estimation methods. Also, we want to investigate the influence of the error propagation of the network ensemble predictions through the PnP(-RANSAC) stage of the pose estimation pipeline. This may be done by using a differentiable PnP implementation like EPro-PnP (Chen et al., 2022) or a deep learning variant like Patch-PnP (Wang et al., 2021).

References

Blundell, C., Cornebise, J., Kavukcuoglu, K., Wierstra, D., 2015. Weight uncertainty in neural network. *ICML*, 37, 1613–1622. 2

Brachmann, E., Michel, F., Krull, A., Yang, M. Y., Gumhold, S., Rother, C., 2016. Uncertainty-Driven 6D Pose Estimation of Objects and Scenes from a Single RGB Image. *2016 IEEE CVPR*, 3364–3372. 3

Chen, H., Wang, P., Wang, F., Tian, W., Xiong, L., Li, H., 2022. EPro-PnP: Generalized End-to-End Probabilistic Perspective-n-Points for Monocular Object Pose Estimation. *2022 IEEE/CVF CVPR*, 2771–2780. 7

Deng, H., Bui, M., Navab, N., Guibas, L., Ilic, S., Birdal, T., 2022. Deep Bingham Networks: Dealing with Uncertainty and Ambiguity in Pose Estimation. *International Journal of Computer Vision*, 130, 1627–1654. 2

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L., 2009. ImageNet: A large-scale hierarchical image database. *2009 IEEE CVPR*, 248–255. 3

Deng, X., Mousavian, A., Xiang, Y., Xia, F., Bretl, T., Fox, D., 2021. PoseRBPF: A Rao–Blackwellized Particle Filter for 6-D Object Pose Tracking. *IEEE Transactions on Robotics*, 37(5), 1328-1342. 2

Drost, B., Ulrich, M., Bergmann, P., Hartinger, P., Steger, C., 2017. Introducing MVTec ITODD - A Dataset for 3D Object Recognition in Industry. *2017 IEEE ICCVW*, 2200–2208. 1

Fort, S., Hu, H., Lakshminarayanan, B., 2020. Deep Ensembles: A Loss Landscape Perspective. *arXiv preprint*. arXiv:1912.02757. 3

Gal, Y., Ghahramani, Z., 2016. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. *ICML*, 48, 1050–1059. 1, 2

Gast, J., Roth, S., 2018. Lightweight Probabilistic Deep Networks. *2018 IEEE/CVF CVPR*, 3369–3378. 2

Gawlikowski, J., Tassi, C. R. N., Ali, M., Lee, J., Humt, M., Feng, J., Kruspe, A., Triebel, R., Jung, P., Roscher, R., Shahzad, M., Yang, W., Bamler, R., Zhu, X. X., 2023. A Survey of Uncertainty in Deep Neural Networks. *Artificial Intelligence Review*, 56, 1513–1589. 3

Gilitschenski, I., Sahoo, R., Schwarting, W., Alexander Amini, S. K., Rus, D., 2020. Deep orientation uncertainty learning based on a bingham loss. *ICLR*. 2

Goodfellow, I. J., Shlens, J., Szegedy, C., 2015. Explaining and Harnessing Adversarial Examples. arXiv:1412.6572. 3

Guo, C., Pleiss, G., Sun, Y., Weinberger, K. Q., 2017. On Calibration of Modern Neural Networks. *ICML*, 70, 1321–1330. 4

Gustafsson, F. K., Danelljan, M., Schon, T. B., 2020. Evaluating Scalable Bayesian Deep Learning Methods for Robust Computer Vision. *2020 IEEE/CVF CVPRW*, 1289–1298. 2

Haugaard, R. L., Buch, A. G., 2022. SurfEmb: Dense and Continuous Correspondence Distributions for Object Pose Estimation with Learnt Surface Embeddings. *2022 IEEE/CVF CVPR*, 6739–6748. 1, 2, 3

Haugaard, R. L., Hagelskjær, F., Iversen, T. M., 2023. Spyro-Pose: SE(3) Pyramids for Object Pose Distribution Estimation. *2023 IEEE/CVF ICCVW*, 2082–2091. 2

He, K., Gkioxari, G., Dollár, P., Girshick, R., 2017. Mask R-CNN. *2017 IEEE ICCV*, 2980–2988. 3

He, K., Zhang, X., Ren, S., Sun, J., 2015. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. *2015 IEEE ICCV*, 1026–1034. 3

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep Residual Learning for Image Recognition. *2016 IEEE CVPR*, 770–778. 3

Heizmann, M., Braun, A., Glitzner, M., Günther, M., Hasna, G., Klüver, C., Krooß, J., Marquardt, E., Overdick, M., Ulrich, M., 2022. Implementing machine learning: chances and challenges. *at - Automatisierungstechnik*, 70(1), 90–101. 1

Hendrycks, D., Gimpel, K., 2018. A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks. *arXiv preprint*. arXiv:1610.02136. 1

Hinton, G., Vinyals, O., Dean, J., 2015. Distilling the Knowledge in a Neural Network. *arXiv preprint*. arXiv:1503.02531. 5

Hodaň, T., Haluza, P., Obdržálek, Š., Matas, J., Lourakis, M., Zabulis, X., 2017. T-LESS: An RGB-D Dataset for 6D Pose Estimation of Texture-less Objects. *2017 IEEE WACV*, 880–888. 2, 5

Hodaň, T., Sundermeyer, M., Drost, B., Labbé, Y., Brachmann, E., Michel, F., Rother, C., Matas, J., 2020. BOP Challenge 2020 on 6D Object Localization. *ECCVW*. 1, 5

Hu, Y., Hugonot, J., Fua, P., Salzmann, M., 2019. Segmentation-Driven 6D Object Pose Estimation. *2019 IEEE/CVF CVPR*, 3380–3389. 2

Huang, W.-L., Hung, C.-Y., Lin, I.-C., 2022. Confidence-Based 6D Object Pose Estimation. *IEEE Transactions on Multimedia*, 24, 3025-3035. 2

Jeon, M.-H., Kim, J., Ryu, J.-H., Kim, A., 2023. Ambiguity-Aware Multi-Object Pose Optimization for Visually-Assisted Robot Manipulation. *IEEE Robotics and Automation Letters*, 8(1), 137–144. 2

Jospin, L. V., Laga, H., Boussaid, F., Buntine, W., Bennamoun, M., 2022. Hands-On Bayesian Neural Networks—A Tutorial for Deep Learning Users. *IEEE Computational Intelligence Magazine*, 17(2), 29–48. 2

Ke, T., Roumeliotis, S. I., 2017. An Efficient Algebraic Solution to the Perspective-Three-Point Problem. *2017 IEEE CVPR*, 4618–4626. 3

Kendall, A., Gal, Y., 2017. What uncertainties do we need in bayesian deep learning for computer vision? *NeurIPS 2017*, 30. 2

Kuleshov, V., Fenner, N., Ermon, S., 2018. Accurate Uncertainties for Deep Learning Using Calibrated Regression. *ICML*, 80, 2796–2804. 4

Labbé, Y., Carpentier, J., Aubry, M., Sivic, J., 2020. CosyPose: Consistent Multi-view Multi-object 6D Pose Estimation. *ECCV*, 574–591. 1, 3

Lakshminarayanan, B., Pritzel, A., Blundell, C., 2017. Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles. *NeurIPS 2017*, 30. 1, 2, 3, 4

Landgraf, S., Wursthorn, K., Hillemann, M., Ulrich, M., 2023. DUDES: Deep Uncertainty Distillation using Ensembles for Semantic Segmentation. *arXiv preprint*. arXiv:2303.09843. 2

Loquercio, A., Segu, M., Scaramuzza, D., 2020. A General Framework for Uncertainty Estimation in Deep Learning. *IEEE Robotics and Automation Letters*, 5(2), 3153–3160. 2

Manhardt, F., Arroyo, D. M., Rupprecht, C., Busam, B., Birdal, T., Navab, N., Tombari, F., 2019. Explaining the Ambiguity of Object Detection and 6D Pose From Visual Data. *2019 IEEE/CVF ICCV*, 6840–6849. 2

McAllister, R., Gal, Y., Kendall, A., van der Wilk, M., Shah, A., Cipolla, R., Weller, A., 2017. Concrete Problems for Autonomous Vehicle Safety: Advantages of Bayesian Deep Learning. *IJCAI-17*, 4745–4753. 1

Mukhoti, J., Kirsch, A., van Amersfoort, J., Torr, P. H., Gal, Y., 2023. Deep Deterministic Uncertainty: A New Simple Baseline. *2023 IEEE/CVF CVPR*, 24384–24394. 2

Okorn, B., Xu, M., Hebert, M., Held, D., 2020. Learning Orientation Distributions for Object Pose Estimation. *2020 IEEE/RSJ IROS*, 10580–10587. 2

Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Nowozin, S., Dillon, J., Lakshminarayanan, B., Snoek, J., 2019. Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift. *NeurIPS 2019*, 32. 1, 2, 3

Park, K., Patten, T., Vincze, M., 2019. Pix2Pose: Pixel-Wise Coordinate Regression of Objects for 6D Pose Estimation. *2019 IEEE/CVF ICCV*, 7667–7676. 1

Peng, S., Liu, Y., Huang, Q., Zhou, X., Bao, H., 2019. PVNet: Pixel-Wise Voting Network for 6DoF Pose Estimation. *2019 IEEE/CVF CVPR*, 4556–4565. 2

Ronneberger, O., Fischer, P., Brox, T., 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, 234–241. 3

Sato, H., Ikeda, T., Nishiwaki, K., 2022. Probabilistic Rotation Representation With an Efficiently Computable Bingham Loss Function and Its Application to Pose Estimation. arXiv:2203.04456. 2

Shi, G., Zhu, Y., Tremblay, J., Birchfield, S., Ramos, F., Anandkumar, A., Zhu, Y., 2021. Fast Uncertainty Quantification for Deep Object Pose Estimation. *2021 IEEE ICRA*, 5200–5207. 2

Sitzmann, V., Martel, J., Bergman, A., Lindell, D., Wetzstein, G., 2020. Implicit Neural Representations with Periodic Activation Functions. *NeurIPS 2020*, 33, 7462–7473. 3

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R., 2014. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research*, 15, 929–1958. 2

Steger, C., Ulrich, M., Wiedemann, C., 2018. *Machine Vision Algorithms and Applications*. 2 edn, Wiley-VCH. 1

Su, Y., Rambach, J., Minaskan, N., Lesur, P., Pagani, A., Stricker, D., 2019. Deep Multi-state Object Pose Estimation for Augmented Reality Assembly. *2019 IEEE ISMAR-Adjunct*, 222–227. 1

Sundermeyer, M., Hodaň, T., Labbé, Y., Wang, G., Brachmann, E., Drost, B., Rother, C., Matas, J., 2023. BOP Challenge 2022 on Detection, Segmentation and Pose Estimation of Specific Rigid Objects. *2023 IEEE/CVF CVPR*, 2784–2793. 1, 3

Thalhammer, S., Bauer, D., Hönig, P., Weibel, J.-B., García-Rodríguez, J., Vincze, M., 2023a. Challenges for Monocular 6D Object Pose Estimation in Robotics. *arXiv preprint*. arXiv:2307.12172. 2

Thalhammer, S., Patten, T., Vincze, M., 2023b. COPE: End-to-end trainable Constant Runtime Object Pose Estimation. *2023 IEEE/CVF WACV*, 2859–2869. 2

Ulrich, M., Hillemann, M., 2024. Uncertainty-Aware Hand-Eye Calibration. *IEEE Transactions on Robotics*, 40, 573–591. 1

Wang, G., Manhardt, F., Tombari, F., Ji, X., 2021. GDR-Net: Geometry-Guided Direct Regression Network for Monocular 6D Object Pose Estimation. *2021 IEEE/CVF CVPR*, 16606–16616. 1, 7

Wursthorn, K., Hillemann, M., Ulrich, M., 2022. Comparison of Uncertainty Quantification Methods for CNN-based Regression. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLIII-B2-2022, 721–728. 2

Xiang, Y., Schmidt, T., Narayanan, V., Fox, D., 2018. PoseCNN: A Convolutional Neural Network for 6D Object Pose Estimation in Cluttered Scenes. *RSS XIV*. 2, 5

Yang, H., Pavone, M., 2023. Object Pose Estimation with Statistical Guarantees: Conformal Keypoint Detection and Geometric Uncertainty Propagation. *2023 IEEE/CVF CVPR*, 8947–8958. 2

Yurtsever, E., Lambert, J., Carballo, A., Takeda, K., 2020. A Survey of Autonomous Driving: Common Practices and Emerging Technologies. *IEEE Access*, 8, 58443–58469. 1

Zhou, G., Gothoskar, N., Wang, L., Tenenbaum, J. B., Gutfreund, D., Lázaro-Gredilla, M., George, D., Mansinghka, V. K., 2023. 3D Neural Embedding Likelihood: Probabilistic Inverse Graphics for Robust 6D Pose Estimation. *2023 IEEE/CVF ICCV*, 21625–21636. 2