Object Pose Estimation with Statistical Guarantees: Conformal Keypoint Detection and Geometric Uncertainty Propagation

Heng Yang and Marco Pavone
NVIDIA Research

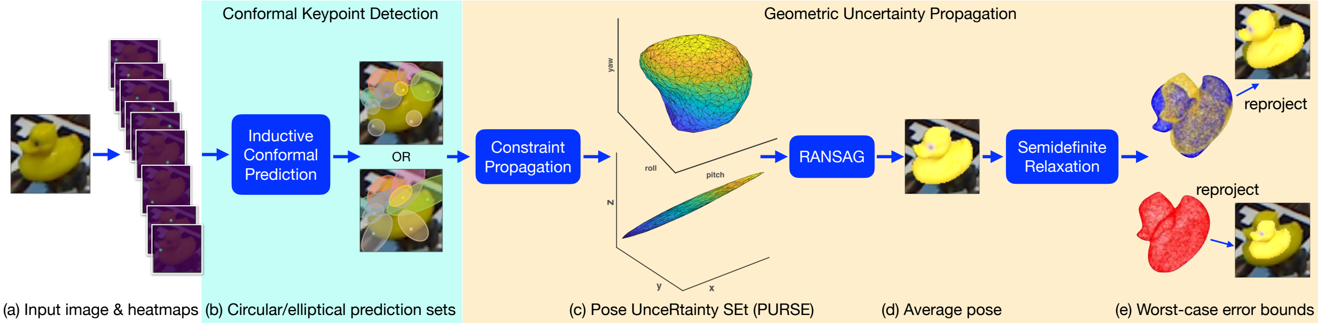
Conformal Keypoint Detection
Geometric Uncertainty Propagation
Inductive Conformal Prediction
OR
Constraint Propagation
yaw
roll
pitch
z
y
x
RANSAG
Semidefinite Relaxation
reproject
reproject
(a) Input image & heatmaps
(b) Circular/elliptical prediction sets
(c) Pose UnceRtainty SEt (PURSE)
(d) Average pose
(e) Worst-case error bounds

Figure 1. Probabilistically correct object pose estimation. Given (a) an input image and heatmap detections of the object semantic keypoints, our method first *conformalizes* the heatmaps into (b) circular or elliptical prediction sets that *guarantee* probabilistic coverage of the groundtruth keypoints (*e.g.*, 90%). Our method then propagates the uncertainty in the keypoints to the object pose, forming (c) a Pose UnceRtainty SEt (PURSE) that contains the groundtruth pose with the same probability. We develop RANdom SAmple averaGing (RANSAG) to sample from PURSE and generate (d) an average pose, and apply semidefinite relaxation to compute (e) worst-case error bounds: the blue duck attains the worst rotation error w.r.t. the (average pose) yellow duck; the red duck attains the worst translation error. Code available: https://github.com/NVlabs/ConformalKeypoint.

Abstract

The two-stage object pose estimation paradigm first detects semantic keypoints on the image and then estimates the 6D pose by minimizing reprojection errors. Despite performing well on standard benchmarks, existing techniques offer no provable guarantees on the quality and uncertainty of the estimation. In this paper, we inject two fundamental changes, namely conformal keypoint detection *and* geometric uncertainty propagation*, into the two-stage paradigm and propose the first pose estimator that endows an estimation with* provable and computable worst-case error bounds*. On one hand, conformal keypoint detection applies the statistical machinery of* inductive conformal prediction *to convert heuristic keypoint detections into circular or elliptical prediction sets that cover the groundtruth keypoints with a user-specified marginal probability* (e.g., 90%)*. Geometric uncertainty propagation, on the other, propagates the geometric constraints on the keypoints to the 6D object pose, leading to a* Pose UnceRtainty SEt (PURSE) *that guarantees coverage of the groundtruth pose with the same probability. The* PURSE*, however, is a nonconvex set that does not directly lead to estimated poses and uncertainties. Therefore, we develop* RANdom SAmple averaGing (RANSAG) *to compute an average pose and apply semidefinite relaxation to upper bound the worst-case errors between the average pose and the groundtruth. On the LineMOD Occlusion dataset we demonstrate: (i) the* PURSE *covers the groundtruth with valid probabilities; (ii) the worst-case error bounds provide correct uncertainty quantification; and (iii) the average pose achieves better or similar accuracy as representative methods based on sparse keypoints.*

1. Introduction

Estimating object poses from images is a fundamental problem in computer vision and finds extensive applications in augmented reality [42], autonomous driving [80], robotic manipulation [60], and space robotics [19]. One of the most popular paradigms for object pose estimation is a *two-stage* pipeline [20, 71, 72, 79, 81, 85, 89, 101], where the first stage detects (semantic) keypoints of the objects on the image, and the second stage computes the object pose by solving an optimization known as *Perspective-n-Points* (PnP) that minimizes reprojection errors of the detected keypoints.

Safety-critical applications call for *provably correct* computer vision algorithms. Existing algorithms in the two-stage paradigm (reviewed in Section 2), however, provide few performance guarantees on the quality of the estimated poses, due to three challenges. (C1) It is difficult to ensure the detected keypoints (typically from neural networks) are close to the groundtruth keypoints. In practice, the first stage often outputs keypoints that are arbitrarily wrong, known as *outliers*. (C2) Robust estimation is employed in the second stage to reject outliers, leading to nonconvex optimizations. Fast heuristics such as RANSAC [26] are widely adopted to find an approximate solution but they cannot guarantee global optimality and often fail without notice. (C3) There is no provably correct *uncertainty quantification* of the estimation, notably, a *formal worst-case error bound* between the estimation and the groundtruth. Though recent work [98] proposed convex relaxations to certify global optimality of RANSAC and addressed (C2), it cannot ensure correct estimation as the optimal pose may be far away from the correct pose when the keypoints are unreliable.

Contributions. We propose a two-stage object pose estimation framework with *statistical guarantees*, illustrated in Fig. 1. Given an input image, we assume a neural network [71] is available to generate *heatmap* predictions of the object keypoints (Fig. 1(a)). Our framework then proceeds in two stages, namely *conformal keypoint detection* (Section 4) and *geometric uncertainty propagation* (Section 5). We first apply the statistical machinery of inductive conformal prediction (introduced in Section 3), with *nonconformity* functions inspired by the design of residual functions in classical geometric vision [39], to conformalize the heatmaps into circular or elliptical prediction sets –one for each keypoint– that guarantee coverage of the groundtruth keypoints with a user-specified *marginal* probability (Fig. 1(b)). This provides a simple and general methodology to bound the keypoint prediction errors (*i.e.*, addressing (C1)). Given the keypoint prediction sets, we reformulate the constraints (enforced by the prediction sets) on the keypoints as constraints on the object pose, leading to a *Pose UnceRtainty SEt* (PURSE) that guarantees coverage of the groundtruth pose with the same probability. Fig. 1(c) plots the boundary of an example PURSE (roll, pitch, raw angles for the rotation, and Euclidean coordinates for the translation). The PURSE, however, is an abstract nonconvex set that does not directly admit estimated poses and uncertainty. Therefore, we develop *RANdom SAmple averaGing* (RANSAG) to compute an average pose (Fig. 1(d)) and employ semidefinite relaxations to upper bound the worst-case rotation and translation errors between the average pose and the groundtruth (Fig. 1(e)). This gives rise to the first kind of *computable* worst-case probabilistic error bounds for object pose estimation (*i.e.*, addressing (C3)). Our PURSE methodology has connections to the framework of *unknown-but-bounded* noise estimation in control theory [63], with special provisions to derive the bounds in a statistically principled way and enable efficient computation.

We test our framework on the LineMOD Occlusion (LM-O) dataset [11] to verify the correctness of the theory (Section 6). First, we empirically show that the PURSE indeed contains the groundtruth pose according to the user-specified probability. Second, we demonstrate the correctness of the worst-case error bounds: when the PURSE contains the groundtruth, our bounds are always larger than, and in many cases close to, the actual errors between the average pose and the groundtruth pose. Third, we benchmark the accuracy of the average pose (coming from RANSAG) with representative two-stage pipelines based on sparse keypoints (*e.g.*, PVNet [72]) and show that the average pose achieves better or similar accuracy.

Limitations. A drawback of our approach, and conformal prediction in general, is that the size of the prediction sets depends on the nonconformity function (whose design can be an art) and may be conservative. Our experiments suggest the bounds are loose when the keypoint prediction sets are large (*e.g.*, giving 180° rotation bound). We discuss challenges and opportunities in tightening the bounds.

2. Related Work

Image-based object pose estimation. We categorize object pose estimation into two paradigms: *single-stage* and *two-stage*. The latter first detects 2D-3D correspondences and then estimates the object pose via solving a PnP problem, while the former produces poses without intermediate correspondences. (i) *Single-stage*. Early methods perform pose estimation via template matching [29, 32, 36]. Recently, deep learning-based approaches such as PoseNet [41] and PoseCNN [95] applied CNNs to directly regress poses. A major challenge of pose regression is the nonlinearity of 3D rotations, and motivated formulating regression as classification [84, 87, 90] or designing better rotation representations [47, 102]. It is also popular to predict multiple pose hypotheses followed by voting [55, 62, 86]. (ii) *Two-stage*. Early research used handcrafted features [50, 56, 78] to establish 2D-3D correspondences and focused on developing algorithms for solving PnP. Notable algorithms include the minimal solver P3P [27, 46] and variants of the nonminimal solver PnP [45, 51, 68, 97]. Outliers (*i.e.*, wrong correspondences) motivated robust estimation based on RANSAC [26], graduated non-convexity [9, 10, 96], branch-and-bound [16, 38, 52], or semidefinite relaxations [98]. Unreliable correspondences soon became the bottleneck and learned correspondences have been predominant. Learned correspondences can be *sparse* or *dense*. Sparse methods define a handful of keypoints and predict locations of the keypoints via direct regression [74, 89], probabilistic heatmap [67, 71], or voting [72]. Dense methods [12, 34, 54, 70, 93, 101] regress for each object pixel the coordinates of its corresponding 3D point. Recent literature focus on end-to-end training via differentiating PnP [13, 15, 20, 21, 37]. Both single-stage and two-stage methods perform well on standard benchmarks [35], but a crucial feature that is missing, especially when deploying computer vision algorithms in safety-critical applications, is that these methods do not provide *provably correct* uncertainty quantification and *formal* error bounds w.r.t. the groundtruth (for either the correspondences or the poses). In this paper, we provide rigorous guarantees by applying conformal prediction to an existing keypoint detection method (the heatmap [71]) and leveraging old and new techniques in computer vision to derive formal error bounds.

Conformal prediction in computer vision. Conformal prediction [92] is a statistical machinery that offers provably correct finite-sample uncertainty quantification without assumptions on the data distribution or the prediction model (*i.e.*, offering a set prediction, instead of a point prediction, that guarantees probabilistic coverage of the groundtruth). *Inductive conformal prediction* [69] is the most popular variant of conformal prediction because it does not require retraining of the prediction models [1, 2, 49]. Applying conformal prediction to computer vision, however, is still in its infancy. Existing works focus on image classification [1, 76], tumor segmentation [3, 8, 94], and bounding box detection [2, 23, 53], which are classification or low-dimensional regression problems. Inspired by these works, our unique contributions in this paper are: (i) we apply conformal prediction to keypoints detection, a high-dimensional regression problem; (ii) we design new nonconformity functions and discuss their connections with classical geometric vision; and (iii) we develop algorithms that propagate the uncertainty after conformal prediction to form prediction sets of 6D poses, which are nonlinear and nonconvex manifold objects.

Performance guarantees. Pose estimation from 2D-2D, 2D-3D, and 3D-3D correspondences are foundational problems in computer vision textbooks [6, 31, 58, 88] and typically boil down to formulating and solving mathematical optimization problems. Benchmarking on simulated and real datasets has been a widely adopted standard for testing different formulations and solvers. However, empirical performance can be misleading without theoretical guarantees. A striking fact is that, though error analysis is an important topic in applied math [17, 24, 43] and control theory [61, 63, 82], there is very limited literature in computer vision that reason about *worst-case estimation errors* between the optimal solution and the groundtruth. A popular heuristic relies on the inverse of the Hessian at an optimal solution, which provides the *Cramer-Rao lower bound* on the covariance of the solution (for linear regression this coincides with the covariance) [88, Section B.6] and thus cannot *upper bound* the estimation errors. Recent works [18, 77, 100] derived error bounds for a few geometric vision problems. However, the bounds either depend on uncheckable assumptions and cannot be computed [77,100], or build on machinery (*e.g.*, sum-of-squares proof [5, 64]) that only applies to estimators based on moment relaxations [18], which are still computationally expensive in practice [99]. In this paper, we develop the first kind of efficiently computable error bounds that only require the assumption of *exchangeability* (which comes from conformal prediction). We justify this assumption on our test dataset and numerically show our bounds can be tight for a subset of the test problems.

3. Inductive Conformal Prediction

Given a set $\{z_i = (x_i, y_i)\}_{i=1}^{l}$ with observation $x_i \in \mathcal{X}$ and label $y_i \in \mathcal{Y}$ such that each $z_i \in \mathcal{Z} := \mathcal{X} \times \mathcal{Y}$ is drawn i.i.d. from an *unknown* distribution on $\mathcal{Z}$, inductive conformal prediction (ICP) provides a *set prediction* $F^{\epsilon}(x) \subseteq \mathcal{Y}$, parameterized by an error rate $0 < \epsilon < 1$, such that given a new sample $z_{l+1} = (x_{l+1}, y_{l+1})$ satisfying an *exchangeability* condition (elaborated in Theorem 1), we have

$$\mathbb{P}\left[y_{l+1} \in F^{\epsilon}(x_{l+1})\right] \geq 1 - \epsilon, \tag{1}$$

i.e., the prediction set F^{ϵ} guarantees to contain the true label y_{l+1} with probability at least $1-\epsilon$.

Training. We start by dividing the dataset into a *proper training set* $\{z_1, \dots, z_m\}$ and a *calibration set* $\{z_{m+1}, \dots, z_l\}$. We shorthand $n = l - m$ as the size of the calibration set. We learn a prediction function $f : \mathcal{X} \to \tilde{\mathcal{Y}}$ from the proper training set using *any* architecture, which allows us to fully exploit the power of modern deep learning. The prediction space $\tilde{\mathcal{Y}}$ can be the same as the label space $\mathcal{Y}$, or can contain auxiliary information such as a heuristic notion of uncertainty (*e.g.*, softmax scores in classification or a heatmap in the case of keypoint detection).

Conformal calibration. We define a *nonconformity* function $S : \mathcal{Z}^m \times \mathcal{Z} \to \mathbb{R}$ to measure how well a given sample $z = (x, y)$ *conforms* to the proper training set. A popular instance of S leverages the learned prediction f:

$$S\left(\{z_1, \dots, z_m\}, (x, y)\right) \overset{e.g.}{=} r(y, f(x)), \tag{2}$$

where $r : \mathcal{Y} \times \tilde{\mathcal{Y}} \to \mathbb{R}$ is a measure of disagreement between the label y and the prediction $f(x)$. For example, consider $\mathcal{Y} = \tilde{\mathcal{Y}} = \mathbb{R}$, one can design $r(y, f(x)) = |y - f(x)|$: if (x, y) poorly conforms to the training set, f will incur large errors. While the function S can be arbitrary (*e.g.*, a learnable neural network [83]), (2) is a convenient definition since f is implicitly dependent on $\{z_i\}_{i=1}^{m}$ and r can incorporate domain-specific knowledge. We then compute the nonconformity scores on the calibration set as $\alpha_i = r(y_i, f(x_i)), i = m+1, \dots, l$, and sort them in *nonincreasing* order $\alpha_{\pi(1)} \geq \dots \geq \alpha_{\pi(n)}$, where $\pi(i) \in \{m+1, \dots, l\}$ is an index permutation.

Conformal prediction. Given a new observation x_{l+1} (with an unknown y_{l+1}) and a user-specified $\epsilon \in (0,1)$, we compute the inductive conformal prediction (ICP) set as

$$F^{\epsilon}(x_{l+1}) = \left\{ y \in \mathcal{Y} \mid \alpha^{y} \leq \alpha_{\pi(\lfloor (n+1)\epsilon \rfloor)} \right\}, \tag{3}$$

where $\alpha^{y} = r(y, f(x_{l+1}))$ is the nonconformity score of the new sample when fixing the true label to be y. In other words, the ICP set (3) outputs the set of all labels that make the nonconformity score of the new sample no greater than $\alpha_{\pi(\lfloor (n+1)\epsilon \rfloor)}$ – the $\lfloor (n+1)\epsilon \rfloor$-th largest nonconformity score in the calibration set. We have the following result stating the probabilistic coverage of the ICP set (3).

Theorem 1 (Validity of ICP Coverage [48, 91, 92]). *If* $z_{m+1}, \dots, z_l$, $z_{l+1} = (x_{l+1}, y_{l+1})$ *are exchangeable,* i.e., *their distribution is invariant under permutation, then*

$$1 - \epsilon \leq \mathbb{P}\left[y_{l+1} \in F^{\epsilon}(x_{l+1})\right] \leq 1 - \epsilon + 1/(n+1) \tag{4}$$

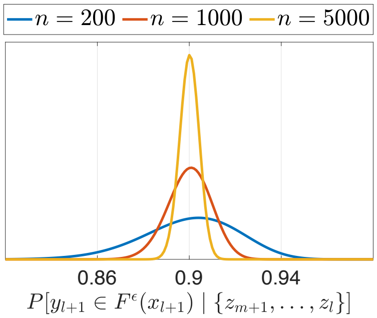

Figure 2. Beta distribution of the conditional coverage in (5) with $\epsilon = 0.1$ and different n. Notice how the conditional probability becomes more concentrated around $1 - \epsilon$ when n increases.

for any $\epsilon \in (0, 1)$. Furthermore, when conditioned on the calibration set, calling $h = \lfloor (n+1)\epsilon \rfloor$, we have

$$\mathbb{P}\left[y_{l+1} \in F^{\epsilon}(x_{l+1}) \mid \{z_{m+1}, \dots, z_l\}\right] \sim \text{Beta}(n+1–h, h). \quad (5)$$

A few remarks are in order about Theorem 1. First, asking $z_{m+1}, \dots, z_l, z_{l+1}$ to be exchangeable is weaker than asking them to be independent. However, this assumption typically fails when the calibration set is a single video sequence, where the image frames $\{z_{m+1}, \dots, z_l\}$ are temporally correlated [57]. Fortunately, as we detail in Section 6, the way the LineMOD Occlusion dataset [11] was collected makes the exchangeability condition easily satisfied, which also suggests best practices to make the exchangeability condition hold in computer vision. Second, the lower bound in (4) can be intuitively proved because under exchangeability, $\alpha_{l+1} := r(y_{l+1}, f(x_{l+1}))$ –the nonconformity score of the new sample with the true label– is *exchangeable* with the nonconformity scores of the calibration samples, and hence *equally likely* to fall in anywhere between the scores $\{\alpha_{\pi(i)}\}_{i=1}^{n}$. Consequently, $\mathbb{P}[y_{l+1} \in F^{\epsilon}(x_{l+1})] = \mathbb{P}\left[\alpha_{l+1} \leq \alpha_{\pi(\lfloor (n+1)\epsilon \rfloor)}\right] = 1 - \lfloor (n+1)\epsilon \rfloor / (n+1) \geq 1 - \epsilon$. The upper bound in (4) states that $1 - \epsilon$ is not overly conservative (indeed tight if n is large). Lastly, the probabilistic guarantee in (4) is *marginal* over the randomness of the calibration set, meaning if one chooses an infinite number of calibration sets, the *average* empirical coverage will converge to $1 - \epsilon$. This, however, implies that the empirical coverage given one calibration set is a random variable that fluctuates as the Beta distribution (5). Fig. 2 plots the Beta distribution at $\epsilon = 0.1$ with different sizes of the calibration set. We observe that as n increases the empirical coverage becomes more concentrated at $1 - \epsilon$. Our experiments show that even with a small ($n = 200$) calibration set, the empirical coverage is close to, and mostly higher than, $1 - \epsilon$.

4. Conformal Keypoint Detection

In this section, we apply the ICP framework in Section 3 to the problem of semantic keypoint detection.

Setup. Denote by $x \in \mathbb{R}^{H \times W \times 3}$ an RGB image picturing an object, by $\boldsymbol{y} = (y_1, \dots, y_K) \in \mathbb{R}^2 \times \dots \times \mathbb{R}^2 := \mathcal{Y}$ the groundtruth locations of K semantic keypoints of the object. We partition a given dataset $\{z_i := (x_i, \boldsymbol{y}_i)\}_{i=1}^{l}$ into a proper training set (of size m) and a calibration set (of size n). We follow the three steps in Section 3 to perform ICP.

Training. We choose the heatmap approach in [71, 79] as the prediction function: given an image x, [79] outputs a set of heatmaps $\boldsymbol{f}(x) = (f(x)_1, \dots, f(x)_K)$, where each $f(x)_k \in \Delta^{HW} := \{v \in \mathbb{R}_{+}^{HW} \mid \sum_{i}^{HW} v_i = 1\}$ predicts the probability distribution of the k-th keypoint lying on each pixel of the image.[1] For convenience, we use $q^j \in \mathbb{R}^2$ to denote the j-th pixel location in x and $f(x)_k^j \in \mathbb{R}_+$ to denote the probability of the k-th keypoint lying on q^j. Let σ_k be the index permutation that sorts $f(x)_k$ in nonincreasing order, *i.e.*, $f(x)_k^{\sigma_k(1)} \geq \dots \geq f(x)_k^{\sigma_k(HW)}$. As we will soon show, choosing the heatmap approach leads to simple and intuitive designs of the nonconformity function.

Conformal calibration. We design the following nonconformity function

$$r(\boldsymbol{y}, \boldsymbol{f}(x)) = \max\{\phi(y_k, f(x)_k)\}_{k=1}^{K} \quad (6)$$

that uses ϕ to score each keypoint and then selects the maximum score. This design considers the worst keypoint detection performance of $\boldsymbol{f}$. We provide two designs of ϕ below.

(a) Peak. Shorthand $p_k = f(x)_k^{\sigma_k(1)}$ as the peak probability in the k-th heatmap and $q_k = q^{\sigma_k(1)}$ as the pixel location attaining the peak probability, we design

$$\phi_{\text{peak}}(y_k, f(x)_k) = p_k \|y_k - q_k\| \quad \text{(peak)}$$

which computes the error between the true keypoint location y_k and the most probable keypoint location q_k and scales the error by the peak probability p_k. ϕ_{peak} describes nonconformity because it becomes larger when the network $\boldsymbol{f}$ is *confidently wrong* (both $\|y_k - q_k\|$ and p_k are large), implying the sample is highly nonconforming.

(b) Covariance. Let $\bar{q}_k = \sum_{j=1}^{J} f(x)_k^{\sigma_k(j)} q^{\sigma_k(j)}$ be the expected location of the top-J most likely detections for the k-th keypoint, and $\Sigma_k = \sum_{j=1}^{J} f(x)_k^{\sigma_k(j)} \cdot (q^{\sigma_k(j)} - \bar{q}_k)(q^{\sigma_k(j)} - \bar{q}_k)^{\mathsf{T}}$ as the covariance, we design

$$\phi_{\text{cov}}(y_k, f(x)_k) = (y_k - \bar{q}_k)^{\mathsf{T}} \Sigma_k^{-1} (y_k - \bar{q}_k) \quad \text{(cov)}$$

which computes the squared Mahalanobis distance [59] from the groundtruth y_k to the top-J keypoint detections (represented by the mean $\bar{q}_k$ and covariance Σ_k).[2] A larger Mahalanobis distance indicates more abnormality of the heatmap $f(x)_k$ (compared to the groundtruth y_k) [28], and hence implies higher nonconformity.

[1] The heatmap in the original paper [71] is not a valid probability distribution as it contains negative values and do not sum up to 1. We remove the negative values and normalize it to be a valid probability distribution.

[2] We only choose the top-J ($J = 100$) most likely detections on the heatmap because the heatmap can be quite noisy in practice.

Using the nonconformity function (6) with (peak) or (cov), we compute the nonconformity scores of the calibration set and sort them as: $\alpha_{\pi(1)} \geq \cdots \geq \alpha_{\pi(n)}$.

Conformal prediction. Given an error rate $\epsilon \in (0,1)$, we first find $\alpha_{\pi(\lfloor (n+1)\epsilon \rfloor)}$. Then, according to the ICP set definition (3) and our nonconformity function (6), we output the ICP set for a new x_{l+1} as

$$\begin{aligned} & F^{\epsilon}(x_{l+1}) \\ = & \{\boldsymbol{y} \in \mathcal{Y} \mid \max\{\phi(y_k, f(x_{l+1}))\}_{k=1}^{K} \leq \alpha_{\pi(\lfloor (n+1)\epsilon \rfloor)}\} \\ = & \quad \{\boldsymbol{y} \in \mathcal{Y} \mid \phi(y_k, f(x_{l+1})) \leq \alpha_{\pi(\lfloor (n+1)\epsilon \rfloor)}, \forall k\}, \end{aligned} \tag{7}$$

where we used $\max\{\phi_1, \ldots, \phi_K\} \leq \alpha$ if and only if $\phi_k \leq \alpha$ for any k. Insert (peak) into (7), we have $F^{\epsilon}_{\text{ball}}(x_{l+1})$ as

$$\left\{ \boldsymbol{y} \in \mathcal{Y} \mid \| y_k - q_{l+1,k} \| \leq \frac{\alpha_{\pi(\lfloor (n+1)\epsilon \rfloor)}}{p_{l+1,k}}, \forall k \right\}, \tag{ball}$$

which defines –for the k-th keypoint– a ball centered at $q_{l+1,k}$ (the most likely detection) with a radius inversely proportional to $p_{l+1,k}$ and proportional to $\alpha_{\pi(\lfloor (n+1)\epsilon \rfloor)}$. Similarly, insert (cov) into (7), we have $F^{\epsilon}_{\text{ellipse}}(x_{l+1})$ as

$$\left\{ \boldsymbol{y} \in \mathcal{Y} \mid (y_k - \bar{q}_{l+1,k})^{\mathsf{T}} \frac{\Sigma^{-1}_{l+1,k}}{\alpha_{\pi(\lfloor (n+1)\epsilon \rfloor)}} (y_k - \bar{q}_{l+1,k}) \leq 1, \forall k \right\}, \tag{ellipse}$$

which defines –for the k-th keypoint– an ellipse centered at $\bar{q}_{l+1,k}$ (the expected location of the top-J detections) with an area proportional to $\det(\Sigma_{l+1,k})$ and $\alpha_{\pi(\lfloor (n+1)\epsilon \rfloor)}$.[3] From (ball) and (ellipse), we observe that the prediction sets become larger when (i) the heatmaps are uncertain, *i.e.*, the peak probability is low or the covariance matrix has large determinant; and (ii) the heatmaps perform poorly on the calibration set, leading to a large $\alpha_{\pi(\lfloor (n+1)\epsilon \rfloor)}$.

Connections to geometric vision. Our nonconformity function bears similarity to the *residual* function in geometric vision [4, 22, 31]. For example, the (peak) and (cov) functions are similar to the (weighted) reprojection error [31], and the "max" in (6) can be connected to seminal work on optimizing the ℓ_∞ norm [39].

Outlier-robust nonconformity? One potential issue of the nonconformity function (6) is that a *single* outlier can inflate the score and the calibration quantile $\alpha_{\pi(\lfloor (n+1)\epsilon \rfloor)}$ and lead to conservative prediction sets (*e.g.*, when $\boldsymbol{f}$ predicts $K-1$ keypoints perfectly but misses one keypoint). A potential remedy in geometric vision is to use robust cost functions [7, 9, 96]. Therefore, a natural question is whether "robustifying" the nonconformity function (6) can lead to better prediction sets. Here we focus on only robustifying ϕ in (6) and provide a negative answer.

Proposition 2 (Invariance of ICP). *Let $\rho : \mathbb{R}_+ \mapsto \mathbb{R}_+$ be any monotonically increasing function. Fixing the calibration set and error rate ϵ, the nonconformity function*

$$r_\rho(\boldsymbol{y}, \boldsymbol{f}(x)) = \max\{\rho(\phi(y_k, f(x)_k))\}_{k=1}^{K} \tag{8}$$

leads to the same ICP set as (6).

The proof of Proposition 2 is presented in Supplementary Material. We conclude that common robust costs, such as ℓ_1, Huber, Geman-McClure, and Barron's adaptive kernel [7, 9] (which are monotonically increasing on $[0, +\infty]$) cannot change the ICP sets by robustifying the individual score ϕ. However, it remains an open question whether changing the "max" operation in (6) can give rise to better ICP sets. For instance, replacing "max" with "$\sum$" in (6) and using the Geman-McClure robust cost $\rho(\phi) = \frac{\phi^2}{1+\phi^2}$ with $\phi = \phi_{\text{peak}}$ results in the following ICP set

$$\left\{ \boldsymbol{y} \in \mathcal{Y} \mid \sum_{k=1}^{K} \frac{p_k^2 \| y_k - q_k \|^2}{1 + p_k^2 \| y_k - q_k \|^2} \leq \alpha_{\pi(\lfloor (n+1)\epsilon \rfloor)} \right\} \tag{9}$$

that does not admit a geometric interpretation that is as simple and intuitive as the (ball) and (ellipse) sets introduced before. In fact, it is indeed the simplicity of (ball) and (ellipse) that enables us to propagate the uncertainty in keypoints to the object pose, as we will show in the next section.

5. Geometric Uncertainty Propagation

Conformalizing the heatmaps gives us prediction sets that guarantee probabilistic coverage of the true keypoints. We unify the prediction sets (ball) and (ellipse) as

$$F^{\epsilon}(x) = \left\{ \boldsymbol{y} \in \mathcal{Y} \mid (y_k - \mu_k)^{\mathsf{T}} \Lambda_k (y_k - \mu_k) \leq 1, \forall k \right\}, \tag{10}$$

where $\mu_k = q_{l+1,k}, \Lambda_k = \frac{p^2_{l+1,k}}{\alpha^2_{\pi(\lfloor (n+1)\epsilon \rfloor)}} \mathbf{I}_2$ for (ball), $\mu_k = \bar{q}_{l+1,k}, \Lambda_k = \frac{\Sigma^{-1}_{l+1,k}}{\alpha_{\pi(\lfloor (n+1)\epsilon \rfloor)}}$ for (ellipse), and we omit the subscript $l+1$ for simplicity.

Why not uncertainty-aware PnP? A popular way to estimate pose from (10) is to solve an uncertainty-aware PnP

$$\begin{aligned} \min_{(R,t) \in \mathrm{SE}(3)} \quad & \sum_{k=1}^{K} (y_k - \mu_k)^{\mathsf{T}} \Lambda_k (y_k - \mu_k) \\ \text{subject to} \quad & y_k = \Pi(RY_k + t), k = 1, \ldots, K \end{aligned} \tag{11}$$

where $Y_k \in \mathbb{R}^3, k = 1, \ldots, K$ are the 3D object keypoints and $\Pi(\cdot)$ denotes the camera projection. We challenge this approach and point out its two drawbacks. First, it is difficult to solve (11) to global optimality due to (i) the nonconvex $\mathrm{SE}(3)$ constraint and (ii) the rational polynomial appearing in $\Pi(\cdot)$. The best known approach to solve (11) relies on either branch-and-bound [68] or local optimization.

[3] The area of $(x - \mu)^{\mathsf{T}} A (x - \mu) \leq 1$ is proportional to $\det(A^{-1})$.

Second, solving (11) typically outputs a *single* optimal pose without uncertainty quantification. Are there other poses that attain similar costs as the optimal pose? How close is the optimal pose to the groundtruth pose? These questions remain not answered in the literature.

Pose UnceRtainty SEt (PURSE). We propose to, instead of solving a PnP problem similar to (11), directly propagate the uncertainty in the ICP sets to the object pose.

Proposition 3 (PURSE). *Let $s_{\text{gt}} = [\text{vec}(R_{\text{gt}})^{\mathsf{T}}; t_{\text{gt}}^{\mathsf{T}}]^{\mathsf{T}}$ be the groundtruth object pose (that lies in front of the camera). Then, the groundtruth keypoints $\boldsymbol{y} = (y_1, \dots, y_K)$ belong to the ICP set $F^{\epsilon}(x)$ in (10) if and only if s_{gt} belongs to the following pose uncertainty set*

$$S^{\epsilon} = \left\{ s \in \text{SE}(3) \;\middle|\; \begin{array}{l} s^{\mathsf{T}} A_k s \leq 0, k = 1, \dots, K \\ b_k^{\mathsf{T}} s > 0, k = 1, \dots, K \end{array} \right\}, \quad \text{(PURSE)}$$

where $A_k \in \mathbb{S}^{12}, b_k \in \mathbb{R}^{12}, k = 1, \dots, K$ are constant matrices dependent on μ_k, Λ_k, Y_k and camera intrinsics.

The detailed proof for Proposition 3 is algebraically involved and postponed to the Supplementary Material. The high-level intuition is, however, straightforward: we plug in $y_k = \Pi(RY_k + t)$ into (10) and obtain K quadratic inequalities of the form $s^{\mathsf{T}} A_k s \leq 0$. The linear inequalities $b_k^{\mathsf{T}} s > 0$ are added to enforce the (transformed) 3D keypoints lie in front of the camera. Proposition 3 implies, if we are $1 - \epsilon$ confident the groundtruth keypoints can be anywhere inside $F^{\epsilon}(x)$, then we should also be confident any pose in (PURSE) can be the groundtruth. Viewing pose estimation as a set estimation with guaranteed probabilistic coverage of the groundtruth is fundamentally different from viewing it as computing a single pose from (11) that is (hopefully) close enough to the groundtruth.

RANdom SAmple averaGing (RANSAG). Verifying if a given pose belongs to the PURSE is straightforward via checking the inequalities in (PURSE). However, the PURSE does not directly give us estimated poses. Therefore, we propose an efficient sampling algorithm called *RANdom SAmple averaGing* (RANSAG) that is analogous to RANSAC [26] and leverages the minimal solver P3P [27], presented in Algorithm 1. The intuition is that, though it is difficult to sample directly in PURSE due to the (nonconvex) constraints, it is easy to sample from the keypoint prediction set (10) due to its simple geometry (balls and ellipses). Thus, at each iteration (line 3) RANSAG samples three keypoints (line 4-5), solves the P3P inverse problem, and accept the poses that belong to the (PURSE) (line 6). RANSAG typically returns around 100 valid samples with $T = 1000$ trials. However, in difficult cases (*e.g.*, when S^{ϵ} is small or even empty) it is possible to obtain zero samples ($S = \emptyset$). In this situation, RANSAG samples $\lfloor T/20 \rfloor$ (default 50) poses without checking if they belong to the PURSE, via sampling K keypoints and solving PnP (line 9-12).[4] After obtaining a set of poses, RANSAG performs rotation averaging (line 14) and translation averaging (line 15) to obtain an average pose $\bar{s}$.[5] Note that RANSAG does not check if $\bar{s}$ lies in the PURSE.

```
Algorithm 1: RANdom SAmple averaGing
1  Input: an ICP set $F^{\epsilon}(x)$ (10) and its corresponding
       (PURSE) $S^{\epsilon}$; maximum trials $T$; initial $\hat{S} = \emptyset$;
2  Output: sample poses $S \subset \text{SE}(3)$ in PURSE, and an
       average pose $\bar{s} \in \text{SE}(3)$;
3  for $\tau \leftarrow 1$ to $T$ do
4      Sample $\{k_1, k_2, k_3\}$ from $[K]$ ($k_1 \neq k_2 \neq k_3$);
5      Sample $\hat{y}_{k_i}, i = 1, 2, 3$ from
           $\{y \in \mathbb{R}^2 \mid (y - \mu_{k_i})^{\mathsf{T}} \Lambda_{k_i} (y - \mu_{k_i}) \leq 1\}$;
6      $\hat{S} \leftarrow \hat{S} \cup (S^{\epsilon} \cap \text{P3P}(\{\hat{y}_{k_i} \leftrightarrow Y_{k_i}\}_{i=1}^{3}))$;
7  end
8  $S = \hat{S}$;
9  if $\hat{S} = \emptyset$ then
10     for $\tau \leftarrow 1$ to $\lfloor T/20 \rfloor$ do
11         Sample $\hat{y}_k, k = 1, \dots, K$ from $F^{\epsilon}(x)$;
12         $\hat{S} \leftarrow \hat{S} \cup \text{PnP}(\{\hat{y}_k \leftrightarrow Y_k\}_{k=1}^{K})$;
13     end
14 end
15 $\bar{R} = \text{proj}_{\text{SO}(3)}(\sum_{(R_j, *) \in \hat{S}} R_j)$;
16 $\bar{t} = \frac{1}{|\hat{S}|} \sum_{(*, t_j) \in \hat{S}} t_j$;
17 return: $S$, $\bar{s} = (\bar{R}, \bar{t})$
```

Worst-case error bounds. To upper bound the errors between the average pose $\bar{s}$ and the groundtruth $(R_{\text{gt}}, t_{\text{gt}})$, we maximize the squared *pose-to-*PURSE distance:

$$d_{\epsilon,\lambda}^2 = \max_{(R,t) \in S^{\epsilon}} \lambda \|R - \bar{R}\|_{\text{F}}^2 + (1 - \lambda) \|t - \bar{t}\|^2 \tag{12}$$

given $\lambda \in [0, 1]$. Particularly, we compute two cases $\lambda = 1$ (the maximum rotation distance) and $\lambda = 0$ (the maximum translation distance). Proposition 3 states the groundtruth $(R_{\text{gt}}, t_{\text{gt}})$ lies in S^{ϵ} with $1 - \epsilon$ probability, hence

$$\|\bar{R} - R_{\text{gt}}\|_{\text{F}} \leq d_{\epsilon,1}, \quad \|\bar{t} - t_{\text{gt}}\| \leq d_{\epsilon,0} \tag{13}$$

holds with probability $1 - \epsilon$.

Computing the bounds. Problem (12) is nonconvex due to the constraints of the (PURSE) S^{ϵ}. We relax the nonconvex problem (12) into a convex semidefinite program (SDP) and employ off-the-shelf solvers to optimize

[4] Here we switch from P3P to PnP because PnP uses all K keypoints and there is less ambiguity in its solution.

[5] In Algorithm 1 we use rotation averaging with the Chordal distance metric. The user is free to choose other single rotation averaging algorithms with different distance metrics [30].

the SDP [14, 40, 98].[6] Two possible outcomes can happen: (i) the optimal SDP value coincides with the optimal value of (12). The relaxation is said to be *exact* and one can extract an optimal solution of (12) from the SDP, or (ii) the relaxation is not exact, but the optimal SDP value still provides an *upper bound* for the optimal value of (12). Therefore, we either exactly compute $d^2_{\epsilon,\lambda}$ or find an upper bound, both can bound the worst-case error (*cf*. (13)).[7]

We end with a remark about computing tighter bounds.

Remark 4 (Best Worst-case Error Bounds)**.** *(12) can be used to bound errors for all possible pose estimators (*e.g.*, from* PnP *(11)). What is the best estimator that attains the smallest error bounds? This boils down to solving*

$$\min_{(\bar{R},\bar{t})\in\mathrm{SE}(3)} \left[\max_{(R,t)\in S^{\epsilon}} \lambda\|R-\bar{R}\|_{\mathrm{F}}^2 + (1-\lambda)\|t-\bar{t}\|^2 \right] \quad (14)$$

whose solution is known as the Chebyshev center *[25, 63] of the* PURSE *S^{ϵ}. Unfortunately, problem (14) is more challenging than (12) and there is no efficient algorithm to solve it to global optimality. In the Supplementary Material, we evaluate the worst-case error bounds for multiple $(\bar{R},\bar{t})$ samples, select the smallest bounds, and compare them with those of the average pose. An interesting future research direction is to explore differentiable optimization [73] or bilevel polynomial optimization [66] to solve (14).*

6. Experiments

We test our approach on the LineMOD Occlusion (LM-O) dataset [11] to (i) justify the exchangeability assumption (Theorem 1) and suggest best practices for applying conformal prediction; (ii) evaluate the empirical coverage of the PURSE and verify the correctness of Theorem 1, and (iii) compute the worst-case error bounds and demonstrate tightness or looseness. We also (iv) show that the average pose achieves better or similar accuracy as other approaches.

Implementation and runtime. We set $T = 1000$ in RANSAG; use OpenGV [44] for P3P and PnP; and add a redundant $\|t\| \leq 5$ in (PURSE) to ensure bounded translation. All procedures are implemented in Python except SDP relaxations are implemented in Matlab. The runtime of RANSAG is comparable to RANSAC and below one second. The runtime of computing (12) via SDPs is around 8 seconds on a workstation with 2.2GHz AMD CPUs. The (second-order) SDP relaxations are almost always exact.

Dataset and exchangeability. The LM-O dataset contains 1214 test images capturing 8 different objects on a table, of which 200 images were chosen by BOP19'20 [35]. We use the 200 images for calibration and the entire 1214 images for testing. As mentioned in Section 3, if the dataset was collected as a single video sequence under natural motion (*e.g.*, a straight line), then the exchangeability assumption would fail. However, [33] described the data collection:

> In order to guarantee a well distributed pose space sampling of the dataset pictures, we *uniformly* divided the upper hemisphere of the objects into *equally distant* pieces and took *at most one image per piece*. As a result, our sequences provide *uniformly distributed views* ...

which indicates the 1214 images are independent (*cf*. [33, Figs. 5-6]) and therefore exchangeable. This demonstrates a good example for data collection –to equally divide the parameter space and collect one observation per division– so the guarantees offered by conformal prediction are valid.

Empirical coverage. Our approach conformalizes the heatmaps [79] as (ball) or (ellipse). The implementation[8] of [79] uses either groundtruth or Faster RCNN [75] bounding boxes, giving four variants of our approach: groundtruth box plus (ball) or (ellipse) (labels: gt-ball, gt-ellipse), and Faster RCNN box plus (ball) or (ellipse) (labels: frcnn-ball, frcnn-ellipse). Fig. 3 left column shows the empirical coverage (*i.e.*, the percentage of images whose groundtruth poses lie in (PURSE)) of all four variants with $\epsilon = 0.1$ and $\epsilon = 0.4$. We see the empirical coverage is around 90% when $\epsilon = 0.1$ and around 60% when $\epsilon = 0.4$, for all 8 objects. Though the empirical coverage can deviate from $1-\epsilon$, it generally stays within ±5% and mostly goes above $1-\epsilon$, which is encouraging given that our calibration set only has size $n = 200$. Fig. 1 (b) plots examples of the prediction sets. More examples are shown in the Supplementary Material.

Worst-case error bounds. Fig. 3 middle column plots the worst-case rotation error bound (x-axis) vs. the actual rotation error between the average pose and the groundtruth (y-axis) for our approach using the gt-ball setup (results for gt-ellipse, frcnn-ball and frcnn-ellipse are similar and provided in the Supplementary Material). First, when the PURSE covers the groundtruth (blue circles), the rotation error bound is always larger than the actual error (*i.e.*, the blue circles never cross the $y = x$ diagonal). Second, when the error rate is increased from $\epsilon = 0.1$ to $\epsilon = 0.4$, we observe a shift of the blue circles towards $y = x$, indicating the error bounds get tightened. Third, our bounds are reasonably tight for most test images (*i.e.*, the bottom-left cluster of blue circles) especially when $\epsilon = 0.4$. However, they can become overly conservative (*i.e.*, the line of blue circles on the right-side boundary) due to the keypoint prediction

[6] We omit the technical details and refer the interested reader to [98, Section 2] for a pragmatic introduction to SDP relaxations. In practice, we use the code provided by [98] in `https://github.com/MIT-SPARK/CertifiablyRobustPerception`, apply a second-order SDP relaxation to (12), and use MOSEK [65] to solve the SDP (in about 8 seconds). Solving a first-order SDP relaxation of (12) takes about 0.1 second but yields looser bounds.

[7] The PURSE can potentially be empty, leading to infeasibility of problem (12). In such cases, empirically the SDP solver returns "`PRIMAL_INFEASIBLE`" (red squares lying on the y-axis of Fig. 3).

[8] `https://github.com/yufu-wang/6D_Pose`

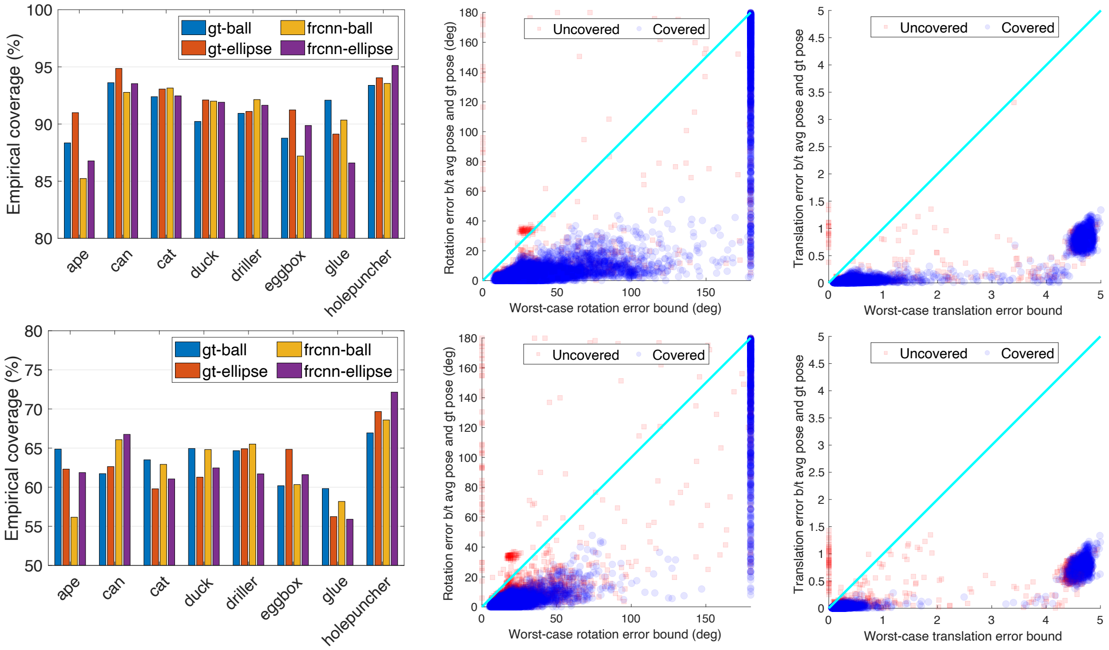

Figure 3. Empirical coverage (left) and worst-case error bounds (middle: rotation, right: translation). Top: $\epsilon = 0.1$, bottom: $\epsilon = 0.4$. For middle and right columns, x-axis represents the worst-case error bounds computed from (12), y-axis represents the actual error between average pose and groundtruth pose. The area below the diagonal $y = x$ indicates correctness of the bounds (*i.e.*, bound $\geq$ error), and points that are closer to the diagonal from below indicate *tighter* bounds (perfect if precisely lie on the diagonal). Blue circles plot cases where the PURSE covers the groundtruth pose and red squares plot cases were the PURSE does not cover the groundtruth. Notice that blue circles never cross the diagonal and our bounds are correct when the PURSE contains the pose (which holds with $1 - \epsilon$ marginal probability).

	Baselines (results adapted from [72])				Conformalized heatmap							
	Tekin	PoseCNN	Oberweger	PVNet	gt-ball		gt-ellipse		frcnn-ball		frcnn-ellipse	
objects	[89]	[95]	[67]	[72]	$\epsilon = 0.1$	$\epsilon = 0.4$	$\epsilon = 0.1$	$\epsilon = 0.4$	$\epsilon = 0.1$	$\epsilon = 0.4$	$\epsilon = 0.1$	$\epsilon = 0.4$
ape	7.01	34.6	69.6	69.14	77.70	79.52	79.26	79.88	70.20	71.01	68.84	69.11
can	11.20	15.10	82.60	86.09	73.41	75.97	75.81	78.13	67.52	69.81	67.69	69.56
cat	3.62	10.40	65.10	65.12	87.36	90.59	89.54	90.11	74.95	80.23	68.98	78.57
duck	5.07	31.80	61.40	61.44	82.71	83.08	84.02	83.55	79.30	80.62	80.06	80.53
driller	1.40	7.40	73.80	73.06	79.32	82.54	81.22	82.04	58.48	65.92	58.06	65.67
eggbox	-	1.90	13.10	8.43	0	0	0.09	0.18	0	0	0	0.14
glue	4.70	13.80	54.90	55.37	56.49	71.08	71.69	72.93	30.03	47.18	41.96	48.26
holepuncher	8.26	23.10	66.40	69.84	81.65	82.89	83.22	84.30	74.96	77.85	76.28	78.18
average	6.16	17.20	60.90	61.06	67.33	70.71	70.61	71.39	56.93	61.58	57.73	61.25

Table 1. Success rates of baseline methods and our conformalized heatmap (using the average pose) based on the 2D projection metric (*i.e.*, a pose estimation is considered successful if the average 2D reprojection error is below 5 pixels).

sets become too large. Fig. 3 right column plots similar results for the translation. The Supplementary Material gives a more detailed analysis of this conservatism, wherein we also solve (12) for multiple samples computed by RANSAG, choose the minimum bound, and compare them with those obtained for the average pose (*cf*. Remark 4).

Accuracy of the average pose. We compare the accuracy of our average pose with other methods according to the 2D projection metric (an estimation is correct if the mean reprojection error is below 5 pixels). Table 1 shows: (i) our average pose achieves significantly better success rates when using groundtruth bounding boxes, and similar success rates when using Faster RCNN; (ii) the accuracy of the average pose increases when ϵ increases.

7. Conclusions

We applied inductive conformal prediction to conformalize heatmap predictions as circular or elliptical prediction sets that guarantee probabilistic coverage of the groundtruth keypoints, propagated the uncertainty in keypoints to the object pose to form a PURSE, designed RANSAG to sample from PURSE and compute an average pose, and used SDP relaxations to bound worst-case estimation errors. We validated our theory on the LineMOD Occlusion dataset. Future research will investigate better nonconformity functions, and applications to other vision problems.

Acknowledgement

We thank Yufu Wang and Kostas Daniilidis for providing the heatmap implementation to detect semantic keypoints; Rachel Luo for discussing the exchangeability assumption in conformal prediction; Luca Carlone for pointing out related work on unknown-but-bounded noise (set membership) estimation in control theory, and anonymous reviewers for providing valuable feedback.

References

[1] Anastasios Angelopoulos, Stephen Bates, Jitendra Malik, and Michael I Jordan. Uncertainty sets for image classifiers using conformal prediction. In *Intl. Conf. on Learning Representations (ICLR)*, 2021. 3

[2] Anastasios N Angelopoulos and Stephen Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511*, 2021. 3

[3] Anastasios N Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. *arXiv preprint arXiv:2208.02814*, 2022. 3

[4] Pasquale Antonante, Vasileios Tzoumas, Heng Yang, and Luca Carlone. Outlier-robust estimation: Hardness, minimally tuned algorithms, and applications. *IEEE Trans. Robotics*, 38(1):281–301, 2021. 5

[5] Boaz Barak and David Steurer. Proofs, beliefs, and algorithms through the lens of sum-of-squares. *Course notes: http://www. sumofsquares. org/public/index. html*, 1, 2016. 3

[6] Timothy D Barfoot. *State estimation for robotics*. Cambridge University Press, 2017. 3

[7] Jonathan T Barron. A general and adaptive robust loss function. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4331–4339, 2019. 5

[8] Stephen Bates, Anastasios Angelopoulos, Lihua Lei, Jitendra Malik, and Michael Jordan. Distribution-free, risk-controlling prediction sets. *Journal of the ACM (JACM)*, 68(6):1–34, 2021. 3

[9] Michael J Black and Anand Rangarajan. On the unification of line processes, outlier rejection, and robust statistics with applications in early vision. *Intl. J. of Computer Vision*, 19(1):57–91, 1996. 2, 5

[10] Andrew Blake and Andrew Zisserman. *Visual reconstruction*. MIT press, 1987. 2

[11] Eric Brachmann, Alexander Krull, Frank Michel, Stefan Gumhold, Jamie Shotton, and Carsten Rother. Learning 6d object pose estimation using 3d object coordinates. In *European Conf. on Computer Vision (ECCV)*, pages 536–551. Springer, 2014. 2, 4, 7

[12] Eric Brachmann, Frank Michel, Alexander Krull, Michael Ying Yang, Stefan Gumhold, et al. Uncertainty-driven 6d pose estimation of objects and scenes from a single rgb image. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 3364–3372, 2016. 2

[13] Eric Brachmann and Carsten Rother. Learning less is more-6d camera localization via 3d surface regression. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 4654–4662, 2018. 2

[14] Jesus Briales, Laurent Kneip, and Javier Gonzalez-Jimenez. A certifiably globally optimal solution to the non-minimal relative pose problem. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 145–154, 2018. 7

[15] Dylan Campbell, Liu Liu, and Stephen Gould. Solving the blind perspective-n-point problem end-to-end with robust differentiable geometric optimization. In *European Conf. on Computer Vision (ECCV)*, pages 244–261. Springer, 2020. 2

[16] Dylan Campbell, Lars Petersson, Laurent Kneip, and Hongdong Li. Globally-optimal inlier set maximisation for simultaneous camera pose and feature correspondence. In *Intl. Conf. on Computer Vision (ICCV)*, pages 1–10, 2017. 2

[17] Emmanuel J Candes, Justin K Romberg, and Terence Tao. Stable signal recovery from incomplete and inaccurate measurements. *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences*, 59(8):1207–1223, 2006. 3

[18] Luca Carlone. Estimation contracts for outlier-robust geometric perception. *arXiv preprint arXiv:2208.10521*, 2022. 3

[19] Bo Chen, Jiewei Cao, Alvaro Parra, and Tat-Jun Chin. Satellite pose estimation with deep landmark regression and nonlinear pose refinement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019. 1

[20] Bo Chen, Alvaro Parra, Jiewei Cao, Nan Li, and Tat-Jun Chin. End-to-end learnable geometric vision by backpropagating pnp optimization. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 8100–8109, 2020. 1, 2

[21] Hansheng Chen, Pichao Wang, Fan Wang, Wei Tian, Lu Xiong, and Hao Li. Epro-pnp: Generalized end-to-end probabilistic perspective-n-points for monocular object pose estimation. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 2781–2790, 2022. 2

[22] Tat-Jun Chin, Zhipeng Cai, and Frank Neumann. Robust fitting in computer vision: Easy or hard? In *European Conf. on Computer Vision (ECCV)*, pages 701–716, 2018. 5

[23] Florence de Grancey, Jean-Luc Adam, Lucian Alecu, Sébastien Gerchinovitz, Franck Mamalet, and David Vigouroux. Object detection with probabilistic guarantees: A conformal prediction approach. In *International Conference on Computer Safety, Reliability, and Security*, pages 316–329. Springer, 2022. 3

[24] Ilias Diakonikolas, Daniel M Kane, Sushrut Karmalkar, Ankit Pensia, and Thanasis Pittas. List-decodable sparse mean estimation via difference-of-pairs filtering. *arXiv preprint arXiv:2206.05245*, 2022. 3

[25] Yonina C Eldar, Amir Beck, and Marc Teboulle. A minimax chebyshev estimator for bounded error estimation. *IEEE transactions on signal processing*, 56(4):1388–1397, 2008. 7

[26] Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981. 1, 2, 6

[27] Xiao-Shan Gao, Xiao-Rong Hou, Jianliang Tang, and Hang-Fei Cheng. Complete solution classification for the perspective-three-point problem. *IEEE Trans. Pattern Anal. Machine Intell.*, 25(8):930–943, 2003. 2, 6

[28] Myung Geun Kim. Multivariate outliers and decompositions of mahalanobis distance. *Communications in statistics-theory and methods*, 29(7):1511–1526, 2000. 4

[29] Chunhui Gu and Xiaofeng Ren. Discriminative mixture-of-templates for viewpoint classification. In *European Conf. on Computer Vision (ECCV)*, pages 408–421. Springer, 2010. 2

[30] Richard Hartley, Jochen Trumpf, Yuchao Dai, and Hongdong Li. Rotation averaging. *International journal of computer vision*, 103:Intl. J. of Computer Vision, 2013. 6

[31] Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003. 3, 5

[32] Stefan Hinterstoisser, Cedric Cagniart, Slobodan Ilic, Peter Sturm, Nassir Navab, Pascal Fua, and Vincent Lepetit. Gradient response maps for real-time detection of textureless objects. *IEEE Trans. Pattern Anal. Machine Intell.*, 34(5):876–888, 2011. 2

[33] Stefan Hinterstoisser, Vincent Lepetit, Slobodan Ilic, Stefan Holzer, Gary Bradski, Kurt Konolige, and Nassir Navab. Model based training, detection and pose estimation of texture-less 3d objects in heavily cluttered scenes. In *Asian conference on computer vision*, pages 548–562. Springer, 2012. 7

[34] Tomas Hodan, Daniel Barath, and Jiri Matas. Epos: Estimating 6d pose of objects with symmetries. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 11703–11712, 2020. 2

[35] Tomas Hodan, Frank Michel, Eric Brachmann, Wadim Kehl, Anders GlentBuch, Dirk Kraft, Bertram Drost, Joel Vidal, Stephan Ihrke, Xenophon Zabulis, et al. Bop: Benchmark for 6d object pose estimation. In *European Conf. on Computer Vision (ECCV)*, pages 19–34, 2018. 2, 7

[36] Daniel P Huttenlocher, Gregory A. Klanderman, and William J Rucklidge. Comparing images using the hausdorff distance. *IEEE Trans. Pattern Anal. Machine Intell.*, 15(9):850–863, 1993. 2

[37] Shun Iwase, Xingyu Liu, Rawal Khirodkar, Rio Yokota, and Kris M Kitani. Repose: Fast 6d object pose refinement via deep texture rendering. In *Intl. Conf. on Computer Vision (ICCV)*, pages 3303–3312, 2021. 2

[38] Yanmei Jiao, Yue Wang, Bo Fu, Qimeng Tan, Lei Chen, Minhang Wang, Shoudong Huang, and Rong Xiong. Globally optimal consensus maximization for robust visual inertial localization in point and line map. In *IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, pages 4631–4638. IEEE, 2020. 2

[39] Fredrik Kahl and Richard Hartley. Multiple-view geometry under the l infinity norm. *IEEE Trans. Pattern Anal. Machine Intell.*, 30(9):1603–1617, 2008. 2, 5

[40] Fredrik Kahl and Didier Henrion. Globally optimal estimates for geometric reconstruction problems. *Intl. J. of Computer Vision*, 74(1):3–15, 2007. 7

[41] Alex Kendall, Matthew Grimes, and Roberto Cipolla. Posenet: A convolutional network for real-time 6-dof camera relocalization. In *Intl. Conf. on Computer Vision (ICCV)*, pages 2938–2946, 2015. 2

[42] Georg Klein and David Murray. Parallel tracking and mapping for small ar workspaces. In *2007 6th IEEE and ACM international symposium on mixed and augmented reality*, pages 225–234. IEEE, 2007. 1

[43] Adam Klivans, Pravesh K Kothari, and Raghu Meka. Efficient algorithms for outlier-robust regression. In *Conference On Learning Theory*, pages 1420–1430. PMLR, 2018. 3

[44] Laurent Kneip and Paul Furgale. Opengv: A unified and generalized approach to real-time calibrated geometric vision. In *IEEE Intl. Conf. on Robotics and Automation (ICRA)*, pages 1–8. IEEE, 2014. 7

[45] Laurent Kneip, Hongdong Li, and Yongduek Seo. Upnp: An optimal o(n) solution to the absolute pose problem with universal applicability. In *European Conf. on Computer Vision (ECCV)*, pages 127–142. Springer, 2014. 2

[46] Laurent Kneip, Davide Scaramuzza, and Roland Siegwart. A novel parametrization of the perspective-three-point problem for a direct computation of absolute camera position and orientation. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 2969–2976. IEEE, 2011. 2

[47] Yann Labbé, Justin Carpentier, Mathieu Aubry, and Josef Sivic. Cosypose: Consistent multi-view multi-object 6d pose estimation. In *European Conf. on Computer Vision (ECCV)*, pages 574–591. Springer, 2020. 2

[48] Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523):1094–1111, 2018. 3

[49] Jing Lei, James Robins, and Larry Wasserman. Distribution-free prediction sets. *Journal of the American Statistical Association*, 108(501):278–287, 2013. 3

[50] Vincent Lepetit, Pascal Fua, et al. Monocular model-based 3d tracking of rigid objects: A survey. *Foundations and Trends in Computer Graphics and Vision*, 1(1):1–89, 2005. 2

[51] Vincent Lepetit, Francesc Moreno-Noguer, and Pascal Fua. Epnp: An accurate o(n) solution to the pnp problem. *Intl. J. of Computer Vision*, 81(2):155–166, 2009. 2

[52] Hongdong Li. Consensus set maximization with guaranteed global optimality for robust geometry estimation. In *Intl. Conf. on Computer Vision (ICCV)*, pages 1074–1080. IEEE, 2009. 2

[53] Shuo Li, Sangdon Park, Xiayan Ji, Insup Lee, and Osbert Bastani. Towards pac multi-object detection and tracking. *arXiv preprint arXiv:2204.07482*, 2022. 3

[54] Zhigang Li, Gu Wang, and Xiangyang Ji. Cdpn: Coordinates-based disentangled pose network for real-time rgb-based 6-dof object pose estimation. In *Intl. Conf. on Computer Vision (ICCV)*, pages 7678–7687, 2019. 2

[55] Joerg Liebelt, Cordelia Schmid, and Klaus Schertler. independent object class detection using 3d feature maps. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8. IEEE, 2008. 2

[56] David G Lowe. Object recognition from local scale-invariant features. In *Intl. Conf. on Computer Vision (ICCV)*, volume 2, pages 1150–1157. Ieee, 1999. 2

[57] Rachel Luo, Shengjia Zhao, Jonathan Kuck, Boris Ivanovic, Silvio Savarese, Edward Schmerling, and Marco Pavone. Sample-efficient safety assurances using conformal prediction. *arXiv preprint arXiv:2109.14082*, 2021. 4

[58] Yi Ma, Stefano Soatto, Jana Košecká, and Shankar Sastry. *An invitation to 3-d vision: from images to geometric models*, volume 26. Springer, 2004. 3

[59] Prasanta Chandra Mahalanobis. On the generalized distance in statistics. National Institute of Science of India, 1936. 4

[60] Lucas Manuelli, Wei Gao, Peter Florence, and Russ Tedrake. kpam: Keypoint affordances for category-level robotic manipulation. In *The International Symposium of Robotics Research*, pages 132–157. Springer, 2019. 1

[61] Maria Cecilia Mazzaro and Mario Sznaier. A set-membership approach to blind identification. In *IEEE Conf. on Decision and Control (CDC)*, volume 5, pages 5176–5181. IEEE, 2004. 3

[62] Frank Michel, Alexander Kirillov, Eric Brachmann, Alexander Krull, Stefan Gumhold, Bogdan Savchynskyy, and Carsten Rother. Global hypothesis generation for 6d object pose estimation. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 462–471, 2017. 2

[63] Mario Milanese and Antonio Vicino. Optimal estimation theory for dynamic systems with set membership uncertainty: An overview. *Automatica*, 27(6):997–1009, 1991. 2, 3, 7

[64] Ankur Moitra. Sum of squares in theoretical computer science. *Sum of Squares: Theory and Applications*, 77:83, 2020. 3

[65] MOSEK ApS. *The MOSEK optimization toolbox for MATLAB manual. Version 8.1.*, 2017. 7

[66] Jiawang Nie, Li Wang, and Jane J Ye. Bilevel polynomial programs and semidefinite relaxation methods. *SIAM Journal on Optimization*, 27(3):1728–1757, 2017. 7

[67] Markus Oberweger, Mahdi Rad, and Vincent Lepetit. Making deep heatmaps robust to partial occlusions for 3d object pose estimation. In *European Conf. on Computer Vision (ECCV)*, pages 119–134, 2018. 2, 8

[68] Carl Olsson, Fredrik Kahl, and Magnus Oskarsson. Optimal estimation of perspective camera pose. In *18th International Conference on Pattern Recognition (ICPR'06)*, volume 2, pages 5–8. IEEE, 2006. 2, 5

[69] Harris Papadopoulos. Inductive conformal prediction: Theory and application to neural networks. In *Tools in artificial intelligence*. Citeseer, 2008. 3

[70] Kiru Park, Timothy Patten, and Markus Vincze. Pix2pose: Pixel-wise coordinate regression of objects for 6d pose estimation. In *Intl. Conf. on Computer Vision (ICCV)*, pages 7668–7677, 2019. 2

[71] Georgios Pavlakos, Xiaowei Zhou, Aaron Chan, Konstantinos G Derpanis, and Kostas Daniilidis. 6-DoF object pose from semantic keypoints. In *IEEE Intl. Conf. on Robotics and Automation (ICRA)*, pages 2011–2018. IEEE, 2017. 1, 2, 4

[72] Sida Peng, Yuan Liu, Qixing Huang, Xiaowei Zhou, and Hujun Bao. Pvnet: Pixel-wise voting network for 6dof pose estimation. *IEEE Trans. Pattern Anal. Machine Intell.*, 2022. 1, 2, 8

[73] Luis Pineda, Taosha Fan, Maurizio Monge, Shobha Venkataraman, Paloma Sodhi, Ricky TQ Chen, Joseph Ortiz, Daniel DeTone, Austin Wang, Stuart Anderson, Jing Dong, Brandon Amos, and Mustafa Mukadam. Theseus: A Library for Differentiable Nonlinear Optimization. In *Advances in Neural Information Processing Systems (NIPS)*, 2022. 7

[74] Mahdi Rad and Vincent Lepetit. Bb8: A scalable, accurate, robust to partial occlusion method for predicting the 3d poses of challenging objects without using depth. In *Proceedings of the IEEE international conference on computer vision*, pages 3828–3836, 2017. 2

[75] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in Neural Information Processing Systems (NIPS)*, 28, 2015. 7

[76] Yaniv Romano, Matteo Sesia, and Emmanuel Candes. Classification with valid and adaptive coverage. In *Advances in Neural Information Processing Systems (NIPS)*, volume 33, pages 3581–3591, 2020. 3

[77] David M Rosen, Luca Carlone, Afonso S Bandeira, and John J Leonard. Se-sync: A certifiably correct algorithm for synchronization over the special euclidean group. *Intl. J. of Robotics Research*, 38(2-3):95–125, 2019. 3

[78] Fred Rothganger, Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce. 3d object modeling and recognition using local affine-invariant image descriptors and multi-view spatial constraints. *Intl. J. of Computer Vision*, 66(3):231–259, 2006. 2

[79] Karl Schmeckpeper, Philip R Osteen, Yufu Wang, Georgios Pavlakos, Kenneth Chaney, Wyatt Jordan, Xiaowei Zhou, Konstantinos G Derpanis, and Kostas Daniilidis. Semantic keypoint-based pose estimation from single RGB frames. *J. of Field Robotics*, 2022. 1, 4, 7

[80] Jingnan Shi, Heng Yang, and Luca Carlone. Optimal pose and shape estimation for category-level 3d object perception. In *Robotics: Science and Systems (RSS)*, 2021. 1

[81] Jingnan Shi, Heng Yang, and Luca Carlone. Optimal and robust category-level perception: Object pose and shape estimation from 2d and 3d semantic keypoints. *arXiv preprint arXiv:2206.12498*, 2022. 1

[82] Torsten Söderström. Errors-in-variables methods in system identification. *Automatica*, 43(6):939–958, 2007. 3

[83] David Stutz, Ali Taylan Cemgil, Arnaud Doucet, et al. Learning optimal conformal classifiers. In *Intl. Conf. on Learning Representations (ICLR)*, 2022. 3

[84] Hao Su, Charles R Qi, Yangyan Li, and Leonidas J Guibas. Render for cnn: Viewpoint estimation in images using cnns trained with rendered 3d model views. In *Intl. Conf. on Computer Vision (ICCV)*, pages 2686–2694, 2015. 2

[85] Jiaming Sun, Zihao Wang, Siyu Zhang, Xingyi He, Hongcheng Zhao, Guofeng Zhang, and Xiaowei Zhou. Onepose: One-shot object pose estimation without cad models. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 6825–6834, 2022. 1

[86] Min Sun, Gary Bradski, Bing-Xin Xu, and Silvio Savarese. Depth-encoded hough voting for joint object detection and shape recovery. In *European Conf. on Computer Vision (ECCV)*, pages 658–671. Springer, 2010. 2

[87] Martin Sundermeyer, Zoltan-Csaba Marton, Maximilian Durner, Manuel Brucker, and Rudolph Triebel. Implicit 3d orientation learning for 6d object detection from rgb images. In *European Conf. on Computer Vision (ECCV)*, pages 699–715, 2018. 2

[88] Richard Szeliski. *Computer vision: algorithms and applications*. Springer Nature, 2022. 3

[89] Bugra Tekin, Sudipta N Sinha, and Pascal Fua. Real-time seamless single shot 6d object pose prediction. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 292–301, 2018. 1, 2, 8

[90] Shubham Tulsiani and Jitendra Malik. Viewpoints and keypoints. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1510–1519, 2015. 2

[91] Vladimir Vovk. Conditional validity of inductive conformal predictors. In *Asian conference on machine learning*, pages 475–490. PMLR, 2012. 3

[92] Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*. Springer Science & Business Media, 2005. 2, 3

[93] Gu Wang, Fabian Manhardt, Federico Tombari, and Xiangyang Ji. Gdr-net: Geometry-guided direct regression network for monocular 6d object pose estimation. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 16611–16621, 2021. 2

[94] Håkan Wieslander, Philip J Harrison, Gabriel Skogberg, Sonya Jackson, Markus Fridén, Johan Karlsson, Ola Spjuth, and Carolina Wählby. Deep learning with conformal prediction for hierarchical analysis of large-scale whole-slide tissue images. *IEEE journal of biomedical and health informatics*, 25(2):371–380, 2020. 3

[95] Yu Xiang, Tanner Schmidt, Venkatraman Narayanan, and Dieter Fox. Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes. In *Robotics: Science and Systems (RSS)*, 2018. 2, 8

[96] Heng Yang, Pasquale Antonante, Vasileios Tzoumas, and Luca Carlone. Graduated non-convexity for robust spatial perception: From non-minimal solvers to global outlier rejection. *IEEE Robotics and Automation Letters*, 5(2):1127–1134, 2020. 2, 5

[97] Heng Yang and Luca Carlone. In perfect shape: Certifiably optimal 3d shape reconstruction from 2d landmarks. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 621–630, 2020. 2

[98] Heng Yang and Luca Carlone. Certifiably optimal outlier-robust geometric perception: Semidefinite relaxations and scalable global optimization. *IEEE Trans. Pattern Anal. Machine Intell.*, 2022. 1, 2, 7

[99] Heng Yang, Ling Liang, Luca Carlone, and Kim-Chuan Toh. An inexact projected gradient method with rounding and lifting by nonlinear programming for solving rank-one semidefinite relaxation of polynomial optimization. *Mathematical Programming*, pages 1–64, 2022. 3

[100] Heng Yang, Jingnan Shi, and Luca Carlone. Teaser: Fast and certifiable point cloud registration. *IEEE Trans. Robotics*, 37(2):314–333, 2020. 3

[101] Sergey Zakharov, Ivan Shugurov, and Slobodan Ilic. Dpod: 6d pose object detector and refiner. In *Intl. Conf. on Computer Vision (ICCV)*, pages 1941–1950, 2019. 1, 2

[102] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 5745–5753, 2019. 2