Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

BOP-Distrib++: Local Symmetry Discovery for 6D Pose Benchmarking Under Visual Ambiguities

Boris Meden

boris.meden@cea.fr

Université Paris-Saclay, CEA, List

Steve Bourgeois

Université Paris-Saclay, CEA, List

Vincent Lepetit

LIGM, Ecole des Ponts, Univ Gustave Eiffel, CNRS

Asma Brazi

Université Paris-Saclay, CEA, List

Fabrice Mayran de Chamisso

Université Paris-Saclay, CEA, List

Research Article

Keywords: Pose estimation, visual ambiguities, benchmark, symmetries

Posted Date: October 8th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10935425/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

BOP-Distrib++: Local Symmetry Discovery for 6D Pose Benchmarking Under Visual Ambiguities

Boris Meden[1*], Steve Bourgeois[1], Vincent Lepetit[2], Asma Brazi[1,2], Fabrice Mayran de Chamisso[1]

[1*]Université Paris-Saclay, CEA, List.
[2]LIGM, Ecole des Ponts, Univ Gustave Eiffel, CNRS, Marne-la-vallée, France.

*Corresponding author(s). E-mail(s): boris.meden@cea.fr;
Contributing authors: steve.bourgeois@cea.fr; vincent.lepetit@enpc.fr; asma.brazi@cea.fr; fabrice.mayran-de-chamisso@cea.fr;

Abstract

6D pose estimation seeks the object pose that best explains a camera observation. While non-ambiguous objects yield unique solutions, ambiguous cases, arising from symmetrical shapes or viewpoints where occlusions hide symmetry-breaking features, result in multi-modal pose distributions. Current benchmarks, however, rely on ground-truth annotations that treat ambiguities as static, global properties. We argue that visual ambiguities must be defined per image to account for viewpoint and partial visibility. Consequently, we first introduce an automatic method to discover an object's local symmetries from its 3D mesh and re-annotate standard datasets with image-wise pose distributions. Second, leveraging this refined ground truth, we re-evaluate state-of-the-art single-pose methods, revealing significant changes in their ranking. Finally, to accommodate recent methods that estimate complete solution sets, we establish a precision/recall evaluation framework, providing the first benchmark for pose distribution methods on real images. https://cea-list.github.io/BOP-Distrib/

Keywords: Pose estimation, visual ambiguities, benchmark, symmetries

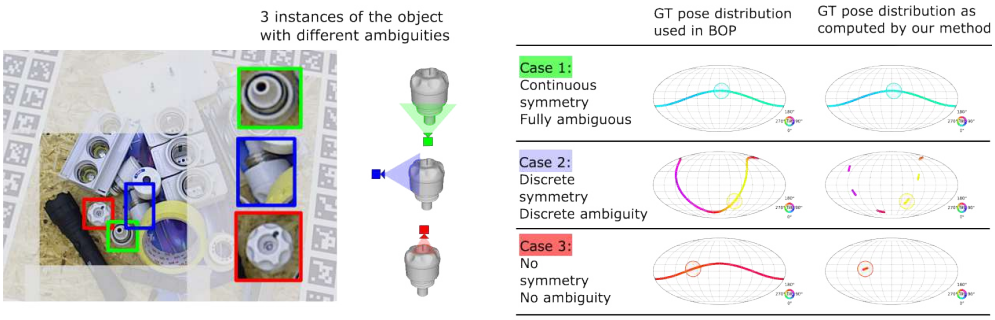

Fig. 1 We provide for the first time 6D pose annotations in the form of a per-image object pose distribution. Current annotations in BOP [1] datasets are given as a single pose, shown here as a circle in the SO(3) representations. BOP also provides a symmetry pattern per object, from which a distribution can be computed (the colored points in SO(3)). Such distribution however does not cover many cases [2]: In this example, when only the core is visible (Case 1), the pose is fully ambiguous and should be represented by a continuous distribution in SO(3). When the sides of the head are visible (Case 2), there are still ambiguities and the distribution is made of 6 modes. When the hole is visible (Case 3), the pose distribution should be concentrated around one non-ambiguous pose. Our method annotates scenes with *per-image* distributions, taking into account the partial occlusions and allowing us to evaluate a predicted pose properly. We show that considering these distributions for evaluation results in a significant change of ranking for the BOP challenge. Such ground truth distributions also become a key asset when it comes to evaluating pose distribution estimation methods [3, 4]. With appropriate metrics, we demonstrate the first quantitative evaluation of pose distribution methods on real images, as an extension to single pose methods.

1 Introduction

Visual 6D pose estimation of an object consists in determining the 3D position and orientation of this object with respect to the camera that explains the observed image. It is a key task in many application domains, such as robotics (*e.g.*grasping or manipulation), augmented reality, industrial quality control, etc. In such contexts, datasets of images with ground truth poses are of primary importance, since they can be used to train learning based methods, but first and foremost to benchmark the performances of proposed approaches. The annotation accuracy of reference benchmarks, such as BOP [5] for 6D pose estimation, is thus crucial since it influences the research directions.

However, annotating images of objects with their 6D poses is a complex task. This is especially true when considered objects include symmetrical parts. Indeed, as illustrated in Figure 1, an image does not necessarily correspond to a single pose solution depending on the viewpoint and occlusions. While symmetrical objects naturally imply multiple solutions, non-symmetrical objects can also yield multiple solutions in case of partial occlusions. Current reference benchmarks [5] ignore these occlusion-induced symmetries, resulting in an imperfect evaluation of the methods. Considering better

ground truth distributions results in a significant change of ranking of pose estimation methods.

In this paper, we propose an automatic method to discover the local symmetries of an object and annotate image-specific pose distributions that account for ambiguities arising from both inherent object symmetries and partial occlusions. Our approach leverages existing annotations, ground-truth visibility masks, and 3D object models to construct a non-parametric pose distribution representation. Unlike approaches based on Bingham distributions [6–8], implicit representations [9, 10], Wigner harmonics [11], matrix Fisher distributions [12] or continuous symmetry groups [13], we do not have a continuous distribution representation. We represent the ground truth pose distributions in a non-parametric way with a set samples of it. This standard, versatile approach captures arbitrary distributions that lack an analytical form, notably in the presence of visual ambiguities, while enabling efficient performance evaluation. Its genericity is demonstrated on T-LESS [14] and YCB-V [15].

We also re-evaluate classical metrics with these annotations to assess the performance of pose estimation methods that return a single pose estimate per image. This yields a strongly modified ranking of current challenger methods.

Moreover, we also consider the few methods that already predict multi-modal pose distributions [3, 4].

Because of the absence of proper annotations, these methods have been evaluated quantitatively only on synthetic images and only qualitatively on real images. Thanks to our annotations, we are able to provide their first evaluation on real data.

In summary, our contributions are the following:

1. A novel automatic method computing a multi-modal 6D pose distribution ground truth from a unique ground truth pose and a proposal set of object symmetries;
2. A comprehensive re-evaluation of 6-DOF single pose estimation methods related to the re-annotation of the T-LESS [14] and YCB-V [15] ground truths;
3. A new evaluation framework of 6-DOF pose distribution estimation methods, which makes it the first multi-modal pose distribution benchmarking on real data.

2 Related Work

Object 6D pose estimation aims at determining an object pose that best explains the camera observation. Multiple equivalent solutions arise when the object is symmetrical or when occlusions (external or self) prevent the observation of symmetry-breaking elements (*e.g.* for a cup [2, 16]).

While the state of the art mostly focused on estimating a single pose [17–24], some recent works focus on estimating the complete set of solutions [2–4, 6–12, 25–28]. As we argue below, to properly evaluate the performance of a method, ground truth annotations including the complete set of solutions for each test image, and metrics that take into account the multiplicity of the solutions, are then required, even for methods that return a single pose per image.

Datasets and 6D pose annotation.

Various techniques have been introduced to annotate image datasets with their 6D poses. For synthetic datasets [29, 30], the ground truth pose is directly available, while, for real datasets, this pose is usually determined with the help of user interaction [31–33], markers [34–36], or a robotic arm [37–39], and refined with ICP in the case of an RGB-D camera [31, 32, 37, 38].

However, such annotations only provide a single pose solution while multiple may exist. For fully symmetric objects, the solutions are always related with the same set of rigid transforms, independently of the viewpoint or the occlusions. This set of transforms, also called *symmetries pattern*, can then be pre-computed offline and applied to the initial solution to recover the whole set of solutions [40].

For non-symmetrical objects, recovering the complete set of solutions is much more challenging since symmetries may arise from the non-visibility of disambiguating parts of the object, those non-visibilities being induced by the viewpoint (self-occlusion), by other elements of the scene (external occlusion), or by the lack of image resolution (resolution occlusion). The transformation set that relates the different solutions is then specific to each image and cannot be pre-computed from the 3D object model.

To our knowledge, no method has been proposed to determine this ground truth transformation set per image. Instead, the current gold standard in 6D pose estimation benchmarks still consists in approximating this per-image transformation set with a unique global transformation set, usually computed with the same method than fully symmetric object with large enough tolerance to ignore small disambiguating elements (*e.g.*surface deviation tolerance of 1.5cm or 10% of the object diameter [40]), and cannot be used to evaluate properly multi-modal pose distribution estimates like [3, 4].

Evaluation metrics.

Performance evaluation usually differs if the method outputs a single pose or a distribution.

For single pose estimation, the accuracy is measured through the deviation of the object surface points when transformed by the estimate and by the ground truth pose (registration error) [16]. Depending on the sensor used—RGB or RGB-D—and the targeted application (*e.g.*robotics or Augmented Reality), deviations can be measured in 3D space (3D distance) or in the image space (reprojection error), and the registration error can be considered as the mean or maximal error over the object surface. In case of multi-valued ground truths due to a symmetrically tagged object, the accuracy corresponds to the minimal error with respect to the set of solutions [15, 16, 30]. In practice, the most commonly used accuracy measures are :

1. Average Distance metric (ADD) [35], and variations for symmetrical objects (ADD-S [15], ADD-H [30]), measure mean Euclidean error between estimated and ground truth surface points, but are replaced by MSSD.
2. Visual Surface Discrepancy (VSD) [32] measures misalignment over the visible surface of the object model. VSD is more expensive to compute than MSSD and

MSPD, and requires a depth image. It is now omitted for new tasks as stated by BOP organizers[1].
3. Maximum Symmetry-Aware Surface Distance (MSSD) [40], similar to ADD, considers the maximal Euclidean error with symmetry management, and provides a 3D error, useful to robotics applications.
4. Maximum Symmetry-Aware Projection Distance (MSPD) [40] measures the maximum reprojection error between projections of the ground truth and estimated surface points, with symmetry management.

Based on these accuracy measures, and inspired from the evaluation of the detection methods, single 6D pose estimation methods are usually evaluated through Precision/Recall. In such evaluation frameworks, an estimated pose is considered as correct if its registration error is below a predefined threshold. The precision corresponds to the rate of correct pose estimations, meaning the ratio of the estimated poses that are correct over the number of estimated poses. The recall corresponds to the rate of correctly registered objects, meaning the ratio between the number of object instances in the dataset whose pose was considered as correct over the number of object instances in the dataset.

Regarding the evaluation of multi-modal 6D pose distribution estimation methods, the accuracy is usually measured as a pose error, meaning the minimal deviation between the inferred rotation and translation and the corresponding nearest ground truth pose [3, 9, 41, 42].

If the method outputs a probability distribution over the whole pose space, the method is usually evaluated through its spread, corresponding to the expectation of the pose error and the log-likelihood between the inferred distribution and the multi-valued solution [3, 9, 10, 41, 42], meaning the sum of the mean log probability at ground truth solutions. While the spread provides a probabilistic measure of the accuracy, the log-likelihood measures how similar the probability distributions are. This penalizes the method if some ground truth solutions are missing in the estimate.

However, not all multi-modal 6D pose estimation methods actually provide a full probability distribution over the whole pose space. Instead, some methods output a set of poses corresponding to some local maxima of the underlying probability distribution [4]. For such methods, performances can be evaluated through the Precision/Recall of the multi-valuated estimation. A pose estimate is then considered as correct with respect to one ground truth pose if its distance in pose space does not exceed a threshold δ. The Precision for a given image is then defined as the ratio of the number of estimated poses whose distance have at least one ground truth pose that does not exceed the threshold δ over the total number of estimated poses. The Recall for an object image is then defined as the ratio between the number of ground truth poses whose distance have at least one estimated pose that does not exceed the threshold δ over the total number of ground truth poses. Similarly to the spread and log likelihood, these precision and recall are related to the multi-modal pose distribution.

Our method combines a ground truth pose and local symmetry patterns to represent the pose distribution for each image. Unlike previous works, the symmetry pattern is adjusted to each image, taking into account its specific viewpoint and

[1] https://bop.felk.cvut.cz/challenges/

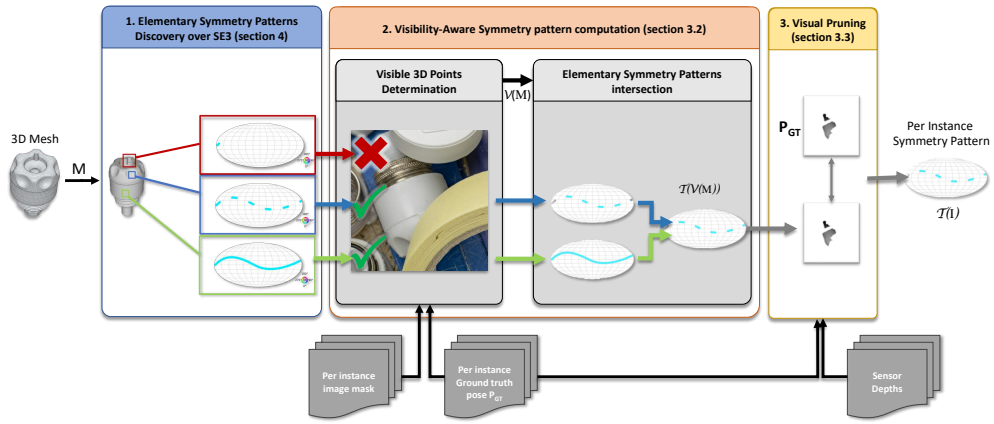

Fig. 2 Overview of the image-specific 6D pose distribution pipeline. (1) Elementary Symmetry Pattern Discovery: Candidate SE(3) symmetry patterns are first extracted from the 3D object model. **(2) Visibility-Aware Symmetry Pattern Computation:** For a given scene image, vertex visibility is evaluated (✓ and ✗ denote visible and occluded vertices, respectively), and a robust intersection of their corresponding symmetry sets is computed. **(3) Visual Pruning :** The intersection is filtered via depth comparison to retain only visually valid symmetries for that specific view. Transforming these valid symmetries by the ground-truth pose yields the multi-modal SE(3) pose distribution of the object instance.

occluded objects (Section 3). Moreover, whereas single pose and multi-modal distribution pose methods are currently evaluated with non-comparable metrics, we introduce new evaluation metrics for multi-modal methods that are homogeneous with single pose methods (Section 5).

3 Method

Consider a 3D object model M observed from a pose P_{GT}. We call symmetry-pattern the set $\mathcal{T} = \{T_i \in \text{SE}(3)\}$ of rigid transformations that, once combined with P_{GT}, generate representations that are geometrically indistinguishable from the one generated by P_{GT}. Depending on the object geometry M and the representation type, the set $\mathcal{T}$ can be limited to the identity transform or can include multiple transformations in addition to the identity.

Previous work well studied the representation corresponding to a 3D surface, or a subset $\mathcal{V}(M) \in M$ of the 3D object surface. In such case, the symmetry pattern $\mathcal{T}(\mathcal{V}(M))$ can be defined through ϵ-symmetries [16]:

$$\epsilon\text{-sym}(\mathcal{V}(M), M) = \{T_i : d(T_i * \mathcal{V}(M), M) < \epsilon\}\,, \tag{1}$$

where ϵ is the maximal deviation tolerance, $*$ represents the application of the transformation T_i onto the point set $\mathcal{V}(M)$ and d measures the max deviation between the sets:
$d(X, M) = \max_{x \in X} ||x - \arg\min_{m \in M}(||x - m||_2)||_2$.

In this paper, we propose to consider the case of representation corresponding to an image I. The symmetry pattern $\mathcal{T}(I)$ associated to the image I corresponds to a subset of the symmetry pattern $\mathcal{T}(\mathcal{V}(M))$ associated to the visible part of the object $\mathcal{V}(M)$ in the image. Indeed, since ϵ-symmetry work in 3D space, they do not take into account the potential self or external occlusions that can be induced by the projection process that produces the image.

Consequently, to estimate $\mathcal{T}(I)$, our solution computes $\mathcal{T}(\mathcal{V}(M))$ then prune the pose visually-inconsistent (Section 4.5). As illustrated in Figure 2, to achieve this estimation efficiently, our solution decomposes the estimation of $\mathcal{T}(\mathcal{V}(M))$ in a two step process: a first pre-computation (Section 3.1) independent of $\mathcal{V}(M)$ and executed only once, and a second step that computes $\mathcal{T}(\mathcal{V}(M))$ (Section 3.2).

3.1 Speeding up the computation of ϵ-sym

Computing $\mathcal{T}(\mathcal{V}(M))$ for each instance in each image of a dataset can quickly become time consuming since it implies to sample the 6D pose space around the ground truth pose. This can explain why such annotation process was never used to our knowledge. We accelerate the ϵ-symmetries computation by introducing terms independent from $\mathcal{V}(M)$ that can be pre-computed.

First, by definition of the ϵ-symmetries, for any subsets $V_1 \subset \mathcal{V}(M)$ and $V_2 \subset \mathcal{V}(M)$, we have:

$$\epsilon\text{-sym}(V_1 \cup V_2, M) = \epsilon\text{-sym}(V_1, M) \cap \epsilon\text{-sym}(V_2, M)\,. \tag{2}$$

Therefore, we can reformulate Equation 1 as:

$$\epsilon\text{-sym}(\mathcal{V}(M), M) = \bigcap_{v_j \in \mathcal{V}(M)} \epsilon\text{-sym}(v_j, M)\,, \tag{3}$$

where v_j is a 3D point and $\epsilon\text{-sym}(v_j, M)$ is the symmetry pattern when the point v_j is the only visible point, which we name *elementary symmetry pattern* of v_j.

With such expression of the ϵ-symmetries, it appears that elementary symmetry patterns can be pre-computed for the whole 3D object model M since they are independent of the image. Only the set of visible points and the intersection of the corresponding pre-computed elementary symmetry patterns need to be computed per image.

3.2 Robust symmetry patterns intersection

In practice, using a strict intersection might reject too many transformations since a transformation T_i would be kept only if it is part of all the elementary symmetry patterns. Instead, we prefer to use a soft intersection, meaning that the transformation T_i will be included even if a few patterns do not include T_i. This soft intersection is achieved by counting, for each transformation T_i, the number of elementary symmetry patterns that include T_i. This counting is defined as follows, in the form of a histogram

over $\mathcal{T}$:

$$\forall i, \mathrm{H}(T_i) \quad = \quad \mathrm{Card}\{v_j \quad | \quad v_j \quad \in \quad \mathcal{V}(M), T_i \quad \in \quad \epsilon\text{-sym}(v_j, M)\} \quad . \quad (4)$$

A strict intersection would only keep the T_i with the maximum value of $\mathrm{H}(T_i)$ which is represented by $\mathrm{H}(Id)$ as the identity of SE(3) and is always in the symmetry pattern. Our soft intersection tolerates a threshold and keeps all the T_i such that $\mathrm{H}(T_i) > \mathrm{H}(Id) - \tau$, where τ represents the minimum size (expressed in number of 3D points) of the symmetry-breaking elements.

4 Elementary Symmetry Patterns Discovery over SE(3)

Discovering the elementary symmetry patterns of an object implies determining, for a dense 3D point cloud of the 3D object surface, the set of rigid transformations $T = (R, t) \in \mathrm{SE}(3)$ that map the point and its local neighborhood onto geometrically consistent regions of the same object. Specifically, it requires to tackle the challenge of revolution surfaces or planar regions that can align continuously with themselves without combinatorial explosion both in memory and processing time.

When a set of initial symmetry candidates is available, such as in the BOP challenge datasets, the remaining degrees of freedom (DoFs) to search over are constrained enough to be tackled via simple sampling strategies. Conversely, in the absence of initial candidates, sampling the unconstrained SE(3) transformation space becomes computationally intractable.

To overcome this challenge, we introduce a two-fold strategy:

1. **Continuous Geometry modeling via Neural SDF:** We encode the object's 3D shape as a continuous, differentiable Signed Distance Field (SDF) using a multi-scale hash-grid architecture [43]. This provides spatial derivatives for direct normal computation and enables non-linear SE(3) pose optimization via differentiable point querying.
2. **Discovering elementary patterns at patch level before point level transfer:** To reduce the memory and compute footprint, we discretize the dense 3D point cloud into a set of overlapping spatial patches. Elementary symmetry patterns are then first generated, refined, and validated at the patch level before being aggregated at 3D point level.

4.1 Implicit and Multi-Scale SDF Representation

To support fast spatial querying and non-linear optimization, we represent the geometry as an implicit Neural Field $\mathcal{S}_\theta : \mathbb{R}^3 \to \mathbb{R}$, parameterized by a multi-resolution hash grid architecture based on RING-NeRF [43] with an SDF output head.

While the hash-grid backbone provides ultra-fast training and query efficiency, applying Random Fourier Feature encoding to the retrieved latent features maps them into a smooth high-frequency space, ensuring continuous differentiability for gradient-based optimization of the SDF. Crucially, the residual grids of RING-NERF

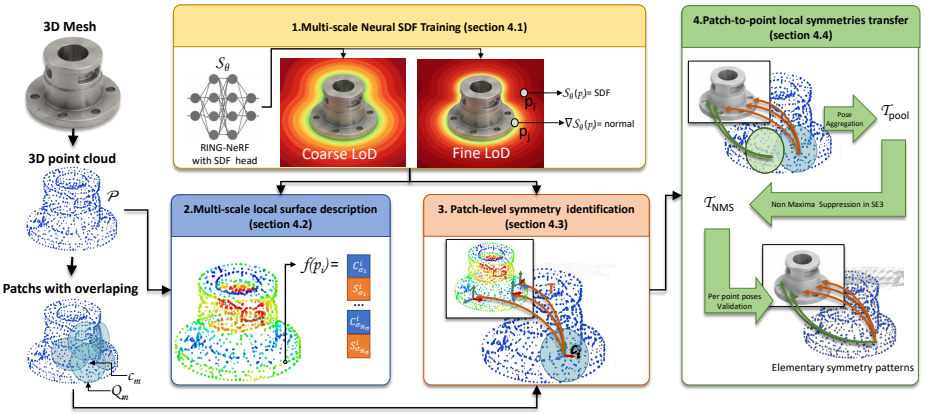

Fig. 3 Overview of the elementary symmetry pattern discovery pipeline. Given a 3D point cloud $\mathcal{P}$ sampled from an object surface, each point p_i is assigned a local surface descriptor $f(p_i)$ and organized into overlapping patches $\mathcal{Q}_m$. For each patch center c_m, corresponding candidate points are identified across $\mathcal{P}$ to generate initial pose hypotheses. These candidate poses are refined through non-linear optimization minimizing the patch-to-surface distance using the differentiable SDF represented by RING-NeRF [43]. Hypotheses with high residual errors are pruned, and the valid poses undergo Non-Maximum Suppression (NMS) over SE(3) to obtain distinct symmetry candidates. Finally, direct point-wise evaluation retains the poses that satisfy local surface and normal alignment criteria, forming the elementary symmetry pattern for each point p_i.

intrinsically represent the object with a continuous Level of Detail (LOD), enabling coarse-to-fine optimization over the distance field and scale-dependent SDF/normal queries.

The network $\mathcal{S}_\theta$ is trained by joint supervision over surface points $\mathcal{P}_{\text{surf}}$ and off-surface spatial samples $\mathcal{P}_{\text{free}}$ using a composite loss function:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{SDF}} + \lambda_{\text{eik}}\mathcal{L}_{\text{eikonal}}, \tag{5}$$

where $\mathcal{L}_{\text{SDF}}$ enforces $\mathcal{S}_\theta(x) = 0$ for $x \in \mathcal{P}_{\text{surf}}$ and supervises the exact SDF value for $x \in \mathcal{P}_{\text{free}}$. To guarantee a valid physical metric space, we impose the Eikonal constraint:

$$\mathcal{L}_{\text{eikonal}} = \sum_{x \in \mathcal{P}_{\text{free}}} \left[\left(\|\nabla_x \mathcal{S}_\theta(x)\|_2 - 1 \right)^2 \right]. \tag{6}$$

To improve the convergence, the SDF is trained with a coarse-to-fine strategy, by gradually activating deeper levels of the multi-resolution grid hierarchy, steadily increasing the level of details of $\mathcal{S}_\theta$.

4.2 Multi-Scale Local Surface Description

The input geometry is sampled into a dense point cloud $\mathcal{P} = \{p_i\}_{i=1}^{N} \subset \mathbb{R}^3$. For each point p_i, we leverage the multi-scale representation of $\mathcal{S}_\theta$ to extract differential surface properties across N_σ different level of details σ.

At each LOD σ, we evaluate the principal curvatures k_1, k_2 ($|k_1| \geq |k_2|$) and their associated principal directions e_1, e_2. From these, we compute the *curvedness* and *shape index* descriptors for a given scale σ as follows:

$$C_\sigma = \sqrt{\frac{k_1^2 + k_2^2}{2}}, \quad S_\sigma = \frac{2}{\pi} \arctan\left(\frac{k_2 + k_1}{k_1 - k_2}\right), \tag{7}$$

where C_σ denotes the *curvedness* (measuring the intensity of surface bending) and $S_\sigma \in [-1, 1]$ represents the *shape index* (capturing local topological identity, *e.g.*cap, rut, ridge, or plane), both for the level of details σ. Since the *shape index* is meaningless when the *curvedness* is null, we associate to each point $p_i \in \mathcal{P}$ the multi-scale local surface descriptor $f(p_i)$ defined as follows:

$$f(p_i) = \left(C^i_{\sigma_1} \cos(\tfrac{S^i_{\sigma_1}\pi}{2}), C^i_{\sigma_1} \sin(\tfrac{S^i_{\sigma_1}\pi}{2}), \ldots, C^i_{\sigma_{N_\sigma}} \cos(\tfrac{S^i_{\sigma_{N_\sigma}}\pi}{2}), C^i_{\sigma_{N_\sigma}} \sin(\tfrac{S^i_{\sigma_{N_\sigma}}\pi}{2})\right) \in \mathbb{R}^{2N_\sigma}. \tag{8}$$

4.3 Patch-level Symmetry Identification

Direct pairwise point matching across all N points incurs an $O(N^2)$ complexity, which is computationally unfeasible. To address this, we partition $\mathcal{P}$ into M overlapping spatial patches $\{\mathcal{Q}_m\}_{m=1}^M$, with $M \ll N$.

4.3.1 Patch-Based Candidate Matching

Patch centers $\mathcal{C} = \{c_m\}_{m=1}^M \subset \mathcal{P}$ are extracted using Farthest Point Sampling. The multi-scale descriptors f of all points are indexed into a GPU-accelerated FAISS [44] structure for fast range queries. For each patch center c_m, target candidate points $\{p_j\}$ are selected if their Mahalanobis distance in the descriptor space satisfies:

$$d_\Sigma(f(c_m), f(p_j)) = \sqrt{(f(c_m) - f(p_j))^\top \Sigma^{-1} (f(c_m) - f(p_j))} < \delta, \tag{9}$$

where $\Sigma \in \mathbb{R}^{2N_\sigma \times 2N_\sigma}$ denotes the covariance matrix computed over the global descriptor distribution, accounting for variance disparities and inter-scale feature correlations.

4.3.2 Pose Hypothesis Generation

While descriptor similarity determines candidate translation alignment, the rotational component of $T = (R, t)$ must be constrained. Aligning the surface normal $n(c_m) = \nabla_x \mathcal{S}_\theta(c_m)$ with the target normal $n(p_j)$ resolves two rotational degrees of freedom (DoF). The remaining rotational DoF around the normal vector depends on local differential surface characteristics:

- **Non-degenerate regions ($|k_1 - k_2| > \epsilon$):** The principal directions (e_1, e_2) are well-defined up to sign. Since the outward normal vector n is fixed and the coordinate frame must remain right-handed ($e_2 = n \times e_1$),

this yields a 2-fold discrete ambiguity. We thus generate 2 candidate rotations corresponding to the π-separated orientations in the tangent plane.

- **Umbilical / Planar regions ($k_1 \approx k_2$):** Since the principal directions are isotropic, the azimuthal orientation is continuously unconstrained. We discretize this continuous 1D search space by sampling N_θ uniform candidate rotations around the normal: $R(\theta_k) \cdot R_{\text{align}}$, where $\theta_k = \frac{2\pi k}{N_\theta}$.

4.3.3 Pose hypothesis Refinement and Validation

By construction, the initial pose hypotheses can be inaccurate and some of them can even be invalid. A process of refinement and validation is then needed.

Each initial candidate pose hypothesis $T^{(0)} \in \text{SE}(3)$ associated with patch $\mathcal{Q}_m$ is refined by minimizing a joint non-linear least-squares objective over the aligned point set. The optimization aligns transformed points with the zero-isosurface while constraining predicted and transformed target normals to be collinear:

$$\min_{T \in \text{SE}(3)} \sum_{p \in \mathcal{Q}_m} \left(\mathcal{S}_\theta(T \cdot p)^2 + \left(1 - \mathbf{n}(T \cdot p)^\top R n_p\right)^2 \right), \tag{10}$$

where $R \in \text{SO}(3)$ is the rotational component of T, n_p is the local patch normal, and $\mathbf{n}(x) = \frac{\nabla_x \mathcal{S}_\theta(x)}{\|\nabla_x \mathcal{S}_\theta(x)\|_2}$ denotes the unit normal field of the implicit surface.

This non-linear optimization is solved using the Levenberg-Marquardt (LM) algorithm. At each LM step, exact SDF values $\mathcal{S}_\theta$ and spatial Jacobians $\nabla_x \mathcal{S}_\theta$ are queried directly from the differentiable RING-NeRF network. By minimizing residuals along directions of high surface curvature while stabilizing geometrically unconstrained directions, such as planar sliding or axial rotations, via $J^\top J$ damping (where J denotes the combined Jacobian of the spatial SDF gradient $\nabla_x \mathcal{S}_\theta$ and normal alignment residuals over $\mathfrak{se}(3)$), the algorithm ensures precise alignment without arbitrary drift along surface invariances. Poses yielding a mean residual error exceeding an adaptive validation threshold are discarded.

4.4 Patch-to-Point Local Symmetries Transfer

Simply assigning a patch's validated poses to all its constituent points causes discretization artifacts: adjacent points sharing similar symmetries receive slightly jittered pose clusters. Consequently, the intersection of those pose distributions, as explained in section 3.2, would result in an empty distribution.

To avoid this issue, the pose distribution of each point is defined among a finite set of pose:

1. We pool the valid pose sets from all patches into a global set $\mathcal{T}_{\text{pool}}$.
2. To eliminate pose redundancies or agglomerations while preserving top-ranked hypotheses, we perform Non-Maximum Suppression (NMS) over SE(3). Candidate poses are sorted by their confidence scores[2] in

[2]The confidence score of a pose is defined as the negative residual error obtained from the preceding patch registration step for this pose.

descending order. A lower-scoring pose T_b is suppressed by a higher-scoring candidate T_a if both its translational distance $d_t(T_a, T_b)$ and rotational angle $d_r(T_a, T_b)$ fall simultaneously below their respective thresholds τ_t and τ_r:

$$(d_t(T_a, T_b) < \tau_t) \quad \text{and} \quad (d_r(T_a, T_b) < \tau_r), \tag{11}$$

where:

$$d_t(T_a, T_b) = \|t_a - t_b\|_2, \quad d_r(T_a, T_b) = \arccos\left(\frac{\mathrm{Tr}(R_a^\top R_b) - 1}{2}\right). \tag{12}$$

This process results in a pruned set of distinct canonical poses $\mathcal{T}_{\text{NMS}} = \{T_k\}_{k=1}^{K}$.

3. Each candidate pose $T_k \in \mathcal{T}_{\text{NMS}}$ is evaluated across all surface points $p_i \in \mathcal{P}$. Each point p_i inherits the subset of validated poses that properly align the point with the SDF.

This yields a consistent, per-point distribution of local rigid symmetries across the entire 3D model.

4.5 Refining ϵ-sym: visual occlusions consideration

Since ϵ-sym does not consider visual occlusions, $\mathcal{T}(\mathcal{V}(M))$ may contain a small set of pose that could induce more visible part of M than $\mathcal{V}(M)$ (see Figure 4).

We follow the VSD metric [16] and determine if a pose T_i should be pruned by rendering the depth image D_{T_i}. For annotation purposes, the minimal disambiguation element size to reject a candidate pose should be the same for all objects. So we remove VSD dependence to object size and count pixels with a depth deviation to ground truth (see Figure 5) greater than δ (mm). If the number of pixels exceeds τ_{pix}, the pose is pruned.

$$Prune\big(D_{GT}, D_{T_i}, V_{GT}, V_{T_i}, \delta\big) = \sum_{p \in V_{GT} \cup V_{T_i}} \begin{cases} 0 & \text{if } p \in V_{GT} \cap V_{T_i} \ \wedge \ \big|D_{GT}(p) - D_{T_i}(p)\big| < \delta \\ 1 & \text{otherwise,} \end{cases} \tag{13}$$

where D_{GT} and V_{GT} are distance image and visible mask of the ground truth, D_{T_i} and V_{T_i} are their T_i counterparts, with visible masks computed from sensor depth.

4.6 Generalization to textured 3D objects

While the process previously described considered only texture-less 3D objects, the method can be extended to textured ones by simply redefining the ϵ-symmetries as:

$$\epsilon\text{-sym}(\mathcal{V}(M), M) = \{T_i : d(T_i * \mathcal{V}(M), M) < \epsilon, \; d_{\text{color}}(T_i * \mathcal{V}(M), M) < \zeta)\}, \tag{14}$$

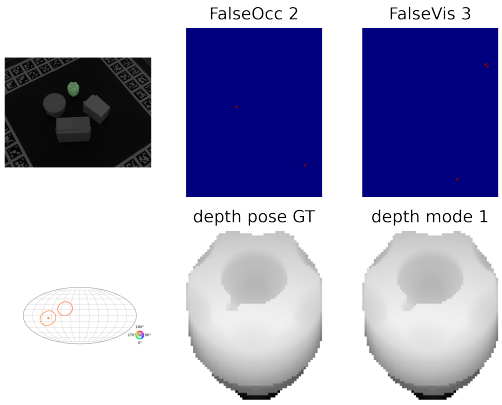

Fig. 4 Depth deviation post-processing analysis. For a given image, we display the depth renderings of the ground truth pose and of one ϵ-sym mode (1 here). They align well.

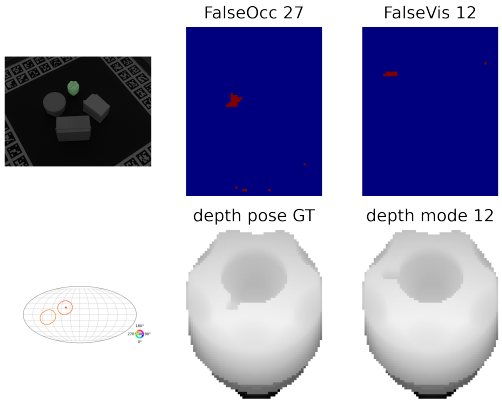

Fig. 5 Depth deviation post-processing analysis. For a given image, we display the depth renderings of the ground truth pose and of one ϵ-sym mode (12). Mode 12 generates several falsely occluded pixels (where the hole should be) and falsely visible pixels (where the hole is but shouldn't be). Mode 12 is rejected by or pruning stage.

where ζ is the color deviation tolerance and with:
$d_{\text{color}}(X, M) = \max_{x \in X} d_{\text{col}}(x, \arg\min_{m \in M}(||x - m||_2))$,
with $d_{\text{col}}(x, m)$ the color distance between vertices x and m.

5 Evaluating 6D Pose Distribution Predictions

Currently, single pose estimation methods and distribution pose estimation methods are evaluated on different datasets (BOP challenge for single poses, SYMSOL for 6D pose distributions) with different metrics (registration error for the first, pose error for the others). With our annotation method that provides the full set of 6D pose solutions even on real data, it becomes possible to evaluate both tasks on the same dataset. However, it would be beneficial if both tasks can be evaluated with comparable metrics.

We therefore propose to keep the evaluation process of single pose method unchanged, but on a more accurate ground truth, and to extend it to pose distributions evaluation. First, whereas pose error is commonly used in 6D pose distribution evaluation process [3, 9, 41], we propose to use the registration error. The latter takes the object 3D shape into account, making the accuracy more meaningful for most applications. It is also the type of error used by the current gold standard benchmark for single pose [5]. Typically, we suggest to use the Maximal Surface Distance (MSD) and Maximum Projective Distance (MPD) metrics. Second, we propose to keep the metrics of Precision and Recall, but we adapt them to measure Precision over the whole estimated distribution of poses instead of a unique pose, and Recall over the whole set of solutions instead of considering a pose to be found if one of its multiple solutions was found.

Precision for pose distribution.

We define the precision in the case of pose distribution prediction as:

$$\mathbf{P_d}(\text{Est}, \text{GT}, \tau_d) = \sum_{P^i_{\text{Est}} \in \text{Est}} p(P^i_{\text{Est}}) \left(\min_{P_{\text{GT}} \in \text{GT}} \mathbf{d}(P^i_{\text{Est}}, P_{\text{GT}}) < \tau_{\mathbf{d}} \right) , \quad (15)$$

where Est and GT are respectively the estimated and the ground truth distributions to compare, and $p(P^i_{\text{Est}})$ represents the probability associated to the i-th element of the estimated distribution [3], $\mathbf{d}$ is the chosen registration distance, and $\tau_{\mathbf{d}}$ the associated threshold.

Recall for pose distribution.

Similarly to the precision, we define the recall in the case of pose distribution prediction as:

$$\mathbf{R_d}(\text{Est}, \text{GT}, \tau_d) = \sum_{P^i_{\text{GT}} \in \text{GT}} \min \left(p(\hat{P}^i_{\text{Est}}), \frac{1}{\text{Card(GT)}} \right) [\mathbf{d}(\hat{P}^i_{\text{Est}}, P^i_{\text{GT}}) < \tau_{\mathbf{d}}],$$

[3] If the method outputs solutions without assigning a probability to each pose, such as [4], a uniform probability over the solutions is used.

$$\text{with } \hat{P}^i_{\text{Est}} = \underset{P_{\text{Est}} \in \text{Est}}{\arg\min}\, \mathbf{d}(P_{\text{Est}}, P^i_{\text{GT}})\,. \quad (16)$$

The probability of $\hat{P}^i_{\text{Est}}$ is clamped to $\frac{1}{\text{Card(GT)}}$ since ground truth poses are considered as equiprobable.

6 Experiments

In this section, after evaluating the performances of our annotation method (Section 6.2), we present and discuss the impact of our ground truth annotations and evaluation metrics onto the evaluation of state-of-the-art solutions for both single pose estimation (Section 6.3) and multi-modal pose distribution estimation (Section 6.4).

Since our annotation process and metrics are related to objects with intrinsic or occlusion-induced symmetries, we perform our evaluations on the T-LESS dataset. Indeed, among the datasets of the BOP challenge [1], T-LESS appears to be the only dataset to exhibit symmetrical objects, with occlusion-induced symmetries, real data and public ground truth annotations: ITODD [31] and HB [45] feature symmetrical objects, with occlusion-induced symmetries, but do not publicly share their ground truths. HOPE [30], Omni6D [46] and IC-BIN [33] have symmetrical shapes, but texture information disambiguates them.

The first experiment focus on a qualitative evaluation of our annotation method (Section 6.2). The second evaluates the impact of our re-annotation of T-LESS and YCB-V datasets on the performance evaluation of the methods of the BOP benchmark [1] on single pose estimation (Section 6.3). The last experiment evaluates for the first time 6D pose distribution methods on a non-synthetic dataset with proper annotations and metrics (Section 6.4).

6.1 Implementation details

In our experiments, we sample uniformly surface points from the CAD models with a resolution of 0.5mm. For the pre-computation of elementary symmetry patterns for a given object, we use a tolerance factor ϵ-sym set to 1mm. When evaluating on the BOP Challenge dataset [1], we leverage the provided per-object symmetry patterns to simplify our elementary symmetry discovery process. Namely, we only execute the final step of Section 4.4, directly instantiating $\mathcal{T}_{\text{NMS}}$ with the official BOP candidate poses.

Regarding the elementary symmetry patterns discovery, we use the following parameters:

- The RING-NeRF hierarchical grid is constituted of 4 levels of grid, with a minimal resolution of 16 and maximal resolution of 256 and a hashmap of $2^2 1$ entries. The latent code is mapped into a high-dimensional feature space using 4 random Fourier frequencies with a scale factor $\sigma = 0.1$. The MLP decoder includes 2 layers of 32 neurons.
- The weight of the eikonal loss is set to $\lambda_{eik} = 1e^{-4}$.

	$\mu_{\|\delta\|}$ (mm)	$\text{med}_{\|\delta\|}$ (mm)	$\max_{\|\delta\|}$ (mm)	$\max_{VSD}$	$\max_{Prune}$ (pix)
From BOP symmetries (sec. 3)	0.218	5.85e-3	0.497	0.061	30
From discovered symmetries (sec. 3 & 4)	0.357	0.090	1.05	0.081	30

Table 1 Quantitative accuracy analysis of our new pose distribution annotations on T-LESS. We report the differences between rendered depths at the ground truth pose and all the points of the corresponding distribution. $\mu_{|\delta|}$, $\text{med}_{|\delta|}$ and $\max_{|\delta|}$ are the mean, median and max of the absolute differences, while $\max_{VSD}$ and $\max_{Prune}$ are max errors of VSD@0.05 and of our pruning, both from BOP symmetries and our complete pipeline with symmetry discoveries.

	$\mathbf{P_{MSD}}$	$\mathbf{R_{MSD}}$	$\mathbf{P_{MPD}}$	$\mathbf{R_{MPD}}$
Symmetry discovery	0.97	0.98	0.84	0.97

Table 2 Quantitative accuracy analysis of our symmetry discovery pipeline on SYMSOL-T. We report the Precision/Recall against analytic symmetries for the per-frame symmetries discovered from whole SE3 space.

- The 3D mesh is sampled 0.5mm spatial resolution and tiled with patches with a radius of 0.1mm spaced every 0.1mm.
- A patch is matched to a point if their mahalanobis distance is lower than $\delta = 2.5$.
- For umbilical or planar regions, the continuous rotation around Z axis is sampled with $N_\theta = 24$ orientations.
- The threshold of the Non Maximal Suppression use a translational threshold $\mathcal{T}_t$ set to 3.125% of the highest dimension of the object, whereas $\mathcal{T}_r = 1$ degree.

Surface visibility $\mathcal{V}(M)$ is computed by Z-buffering using ground truth pose P_{GT} and the 3D model of the object. For the robust symmetry pattern intersection, the soft intersection tolerance factor τ was experimentally adjusted to 28 3D points, resulting in a minimal disambiguating size of roughly 2.5×2.5mm^2, $\delta = 5$mm and $\tau_{pix} = 30$pix.

6.2 Validation of the method

Figure 7 illustrates the coherence between the distribution and the visibility of the disambiguating elements. The ground truth translation is always reduced to a single 3D point as T-LESS exhibit no translational ambiguity.

Regarding quantitative evaluation of the symmetry pattern accuracy, Table 1 reports the max deviation error (in mm) between the depth image rendered from the GT pose P_{GT} and the ones rendered from the combination of P_{GT} with any pose of the symmetry pattern. We observe that the rendered depths are really close.

We resort to SYMSOL-T [4, 9] to quantify the effectiveness of our Elementary Symmetry Patterns Discovery process over SE(3) presented in section 4. Unlike generic 3D models, the symmetry patterns are analytically accessible for these geometric shapes. We compare the results of our discovery method to these analytic ground truth, using our precision/recall metrics over pose distributions.

	BOP Contenders	BOP [5] (object-wise annotations) MSSD	MSPD	Mean	Rank	BOP-Distrib (our image-wise annotations) MSSD	MSPD	Mean	Loss	Rank
	T-LESS [14] **BOP Challenge 2023 [1]: pose estimation of unseen objects**									
RGB	foundpose [47]	55.0	62.3	58.6	1	45.5	52.1	48.8	-9.8	3(↓-2)
RGB	gigaposemegapose-5-hypothesis [48, 49]	54.3	62.0	58.1	2	45.3	52.3	48.8	-9.3	2
RGB	gigaposemegapose-1-hypothesis [48, 49]	51.5	59.2	55.3	3	38.5	45.1	41.8	-13.5	9(↓-6)
RGB	genflow-multihypo16-rgb [50]	50.9	57.9	54.4	4	45.9	52.4	49.1	-5.3	1(↑3)
RGB	genflow-multihypo-rgb [50]	50.4	57.3	53.9	5	44.9	51.4	48.1	-5.8	4(↑1)
RGB	genflow-multihypo16 [50]	52.8	54.8	53.8	6	46.7	48.6	47.7	-6.1	5(↑1)
RGB	genflow-multihypo [50]	51.6	53.6	52.6	7	45.2	47.2	46.2	-6.4	6(↑1)
RGB	foundpose [47]	49.1	55.6	52.4	8	37.7	43.4	40.6	-11.8	11(↓-3)
RGB	foundpose [47]	49.1	55.6	52.3	9	37.7	43.4	40.5	-11.8	12(↓-3)
RGB	cnos-fastsammegapose-multihyp-10 [49, 51]	48.5	55.9	52.2	10	40.6	47.1	43.8	-8.4	7(↑3)
RGB	cnos-fastsammegapose-multihyp [49, 51]	48.4	55.6	52.0	11	40.5	47.1	43.8	-8.2	8(↑3)
RGB	cnos-fastsammegapose-multihyp-teaserpp [49, 51]	48.4	51.3	49.9	12	40.3	43.0	41.7	-8.2	10(↑2)
RGB-D	foundationpose [52]	62.9	64.1	63.5	1	56.7	58.0	57.4	-6.1	1
RGB-D	gigaposegenflowkabsch [48]	56.3	58.6	57.5	2	49.4	51.7	50.6	-6.9	2
RGB-D	sam6d [53]	52.9	54.5	53.7	3	39.7	41.2	40.4	-13.3	3
RGB-D	sam6d [53]	52.1	53.5	52.8	4	38.3	39.7	39.0	-13.8	5(↓-1)
RGB-D	sam6d-fastsam [53]	49.9	51.1	50.5	5	37.9	39.1	38.5	-12.0	6(↓-1)
RGB-D	sam6d-cnosmask [53]	50.0	50.9	50.4	6	39.3	40.3	39.8	-10.6	4(↑2)
RGB-D	sam6d-cnosfastsam [53]	49.3	50.5	49.9	7	37.2	38.4	37.8	-12.1	7
RGB-D	sam6d-zeropose [53]	44.4	45.6	45.0	8	33.5	34.8	34.2	-10.8	8
RGB-D	zeropose-multi-hypo-refinement-defaultseg [54]	41.5	42.0	41.8	9	29.4	30.1	29.7	-12.1	9
	T-LESS [14] **BOP Challenge 2022 [5]: pose estimation**									
RGB	hccepose-2024-pbr	82.9	95.8	89.4	1	53.8	65.2	59.5	-29.9	5(↓-4)
RGB	hcceposeefficientnet-b4-default-2d-bbox	81.4	94.8	88.1	2	53.1	64.5	58.8	-29.3	6(↓-4)
RGB	cosypose-eccv20-syntreal-8views [20]	83.6	90.7	87.2	3	57.2	64.0	60.6	-26.6	4(↓-1)
RGB	hccepose-default-2d-bbox	79.2	93.2	86.2	4	51.4	63.1	57.2	-29.0	8(↓-4)
RGB	gdrnppdet-pbrrealgenflow-multihypo-rgb [22, 55]	78.4	92.2	85.3	5	69.9	82.9	76.4	-8.9	2(↑3)
RGB	leroy-fuseocclu-rgb	78.0	91.7	84.8	6	70.4	83.3	76.9	-7.9	1(↑5)
RGB	gpose2023-rgb [56]	76.6	92.9	84.7	7	50.2	64.0	57.1	-27.6	9(↓-2)
RGB	gdrnpp-pbr-rgb-mmodel [22, 55]	76.3	92.4	84.4	8	50.0	63.6	56.8	-27.6	10(↓-2)
RGB	gdrnpp-pbrreal-rgb-mmodel [22, 55]	76.0	91.3	83.6	9	48.8	61.7	55.2	-28.4	13(↓-4)
RGB	cosypose-eccv20-syntreal-4views [20]	79.5	86.4	83.0	10	52.7	59.3	56.0	-27.0	12(↓-2)
RGB	mrc-net [57]	78.5	87.1	82.8	11	52.3	60.3	56.3	-26.5	11
RGB	gdrnppdet-pbrrealmegapose-multihyp [22, 55]	74.0	87.4	80.7	12	61.6	73.8	67.7	-13.0	3(↑9)
RGB	sc6d [58]	72.1	85.3	78.7	13	52.6	64.1	58.3	-20.4	7(↑6)
RGB-D	gpose2023 [56]	92.1	94.6	93.4	1	86.5	89.1	87.8	-5.6	1
RGB-D	gdrnppv2-rgbd-pbrreal [22, 55]	90.2	92.6	91.4	2	84.8	87.2	86.0	-5.4	3(↓-1)
RGB-D	gpose2023-pbr [56]	90.3	92.6	91.4	3	84.7	87.2	86.0	-5.4	2(↑1)
RGB-D	modalocclusion-rgbd	89.7	92.9	91.3	4	83.2	86.4	84.8	-6.5	4
RGB-D	hcceposebf-2024-ref	89.0	91.7	90.3	5	71.7	74.7	73.2	-17.1	6(↓-1)
RGB-D	gdrbpp-pbrreal-rgbd-mmodel-fast **v1.4** [22, 55]	88.7	91.5	90.1	6	58.1	61.9	60.0	-30.1	13(↓-7)
RGB-D	gdrnpp-pbrreal-rgbd-mmodel **v1.3** [22, 55]	88.4	90.9	89.6	7	60.6	63.7	62.2	-27.4	10(↓-3)
RGB-D	gdrnpp-pbrreal-rgbd-smodel **v1.2** [22, 55]	87.1	90.3	88.7	8	57.5	61.2	59.3	-29.4	15(↓-7)
RGB-D	zebraposesat-effnetb4-refineddefdet-2023 [24]	87.3	90.2	88.7	9	60.2	63.4	61.8	-26.9	11(↓-2)
RGB-D	gdrnpp-pbr-rgbd-mmodel **v1.1** [22, 55]	87.2	90.1	88.7	10	60.8	64.2	62.5	-26.2	9(↑1)
RGB-D	hipose-cvpr24 [59]	86.6	88.8	87.7	11	58.0	60.8	59.4	-28.3	14(↓-3)
RGB-D	defaultdetection-pfa-mixpbr-rgb-d	85.5	87.7	86.6	12	59.2	62.0	60.6	-26.0	12
RGB-D	zebraposesat-effnetb4defdet-2023 [24]	79.1	93.0	86.1	13	51.8	63.4	57.6	-28.5	17(↓-4)
RGB-D	zebraposesat-effnetb4pbr-only-defdet-2023	77.7	92.0	84.9	14	50.7	62.8	56.8	-28.1	18(↓-4)
RGB-D	gdpnpp-pbrreal-rgbd-mmodel-officialdet **v1.0** [22, 55]	83.4	85.6	84.5	15	57.5	60.2	58.9	-25.6	16(↓-1)
RGB-D	surfemb-pbr-rgbd-lin [23]	82.9	85.9	84.4	16	75.8	78.7	77.2	-7.2	5(↑11)
RGB-D	poseio	81.7	86.0	83.8	17	69.3	73.6	71.4	-12.4	8(↑9)
	YCB-V [15] **BOP Challenge 2022 [5]: pose estimation**									
RGB	cosypose-eccv20-syntreal-8views [20]	88.5	88.0	88.2	1	83.3	82.4	82.9	-5.3	2(↓-1)
RGB	cosypose-eccv20-syntreal-4views [20]	86.9	86.6	86.8	2	82.1	81.3	81.7	-5.1	4(↓-2)
RGB	mrpe-pbrreal-rgb-smodel	85.7	87.3	86.5	3	83.1	84.4	83.7	-2.8	1(↑2)
RGB	magic	85.9	85.4	85.6	4	82.9	82.3	82.6	-3.0	3(↑1)
RGB-D	gdrnppv2-rgbd-pbrreal [22, 55]	96.3	92.7	94.5	1	94.2	90.2	92.2	-2.3	1
RGB-D	gpose2023 [56]	96.2	92.6	94.4	2	94.2	90.2	92.2	-2.2	2
RGB-D	gdrnpp-pbrreal-rgbd-mmodel [22, 55]	95.7	91.9	93.8	3	90.1	86.2	88.1	-5.7	3
RGB-D	dsgc6d	93.0	90.1	91.5	4	87.5	85.2	86.4	-5.1	4
	T-LESS [14] **Best mode of pose distribution methods**									
RGB	LiePose Diffusion [4]	60.1	92.2	76.1	1	48.3	76.1	62.2	-13.9	2(↓-1)
RGB	SpyroPose [3]	61.2	75.3	68.3	2	57.8	71.6	64.7	-3.6	1(↑1)

Table 3 Results of BOP 2023 on T-LESS [14] and YCB-V [15] with our annotations. We report the top contenders results of the **pose estimation of unseen objects challenge** and the **pose estimation challenge** evaluated with **MSSD** and **MSPD** on both the official ground truth (object-wise) and our ground truth (image-wise). We rank the methods with the mean of **MSSD** and **MSPD** (similar to BOP, except we exclude **VSD** as BOP starts to abandon it) and re-rank them with the mean of **MSSD** and **MSPD** computed on our ground truth distributions. The drop in the scores highlighted by the **Loss** column ($\textbf{Loss} = \textbf{Mean}_{\textbf{BOP-Distrib}} - \textbf{Mean}_{\textbf{BOP}}$) produces drastic changes in the rankings. We also show the same scores for the **best mode of distribution estimation methods**, however, both of them use ground truth bounding boxes instead of a detector, so direct comparison would be unfair. This highlights the importance to consider image level symmetries as we proposed, to evaluate accurately pose estimation methods. We also processed YCB-V. The **Loss** is much lower as there are fewer symmetrical objects. Names of the methods are automatically processed from BOP csv result files.

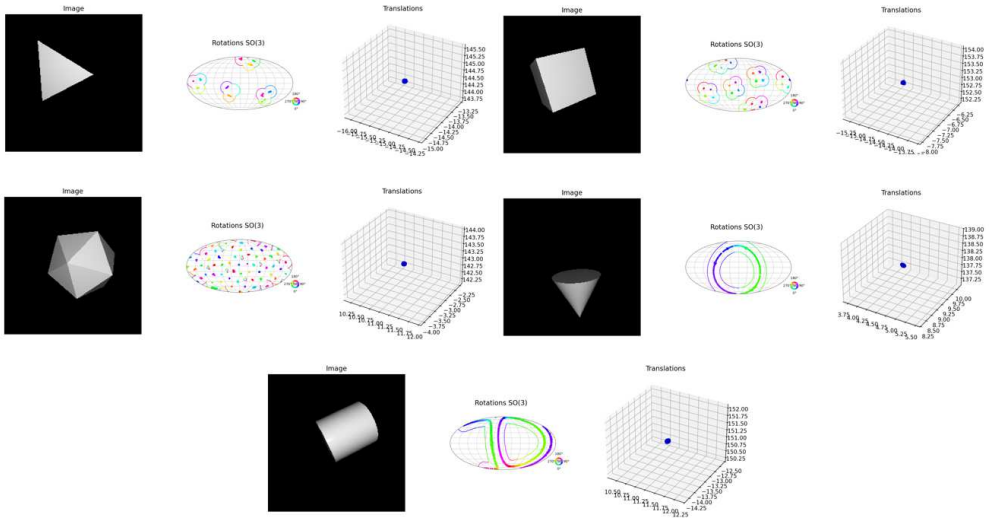

Fig. 6 Visualizations of our symmetry discovery against analytic symmetries on SYMSOL-T. We display SE(3) ground truth distributions for all shapes of SYMSOL-T [4, 9]. Circle on orientation diagram represents the unique ground truth pose provided as input to our method. Colors link objects to their distributions.



Fig. 7 Visualizations of our ground truth. We display SE(3) ground truth distributions for scene 1 of T-LESS [14]. Circle on orientation diagram represents the unique ground truth pose provided as input to our method. Colors link objects to their distributions.

6.3 Single Pose Estimation Evaluation

Since May 2023, raw results of BOP Challenge [1] submissions have become publicly available to allow in-depth analysis. This experiment reprocess these results to compare the performance evaluations of state-of-the-art methods for both the original 6D annotations and our annotations. We considered top contenders of the BOP challenge for the pose estimation and pose estimation for unseen objects tasks. We also included two methods for the multi-modal pose distribution estimation task [3, 4] for which we kept only the highest confidence mode.

Experimental protocol.

We evaluate the current (July 2024) top contenders of the BOP Challenge on the T-LESS dataset [14] on the **pose estimation** task and the **pose estimation of unseen object** task together with [3, 4]. We conduct a similar evaluation with YCB-V [15].

This evaluation is based directly on the raw pose results provided by the contenders and the official evaluation scripts of BOP challenges, both being publicly available on the Challenge website. Our method ranking is based on the mean of Recalls computed on MSPD and MSSD accuracies.

Results.

Table 3 reports the results. We can observe a large impact of the ground truth on the methods results (see **Loss** column) and ranking. In terms of ranking, the well established **pose estimation** task faces the biggest changes, with gdrnppdet-pbrrealmegapose-multihyp moving from the 12th to the 3rd position for RGB methods and surfemb-pbr-rgbd-lin [23] moving from the 16th to the 5th place for the RGB-D methods. For **pose estimation of unseen objects**, the difference of ranking can reach 6 places.

We observe large impacts on the recall performances as well. Whereas 29 of the 30 contenders of **pose estimation** task have a mean recall up to 80% and even up to 90% for 6 methods with the original annotations, only 4 methods exceed this score with our annotations, with a majority of methods having less than 65%. Similarly, most of the contenders for the **pose estimation of unseen objects** task have a mean recall that exceeds 50% with the original annotations whereas only the 2 best RGB-D methods exceed this score once evaluated with our annotations. As shown in Figure 1 (Case 3), the image-wise annotation rejects poses when the image is not ambiguous. The Loss is explained by inaccurate pose estimates that were mistakenly validated by previous object-wise annotations (see supplementary materials for visualizations).

6.4 Pose Distribution Evaluation

We provide now the first evaluation of 6D pose distribution estimation methods on a dataset of real images.

Experimental protocol.

Similarly to Section 6.3, the methods were trained on the train set of T-LESS. The ground truth used for this dataset is the same than the one of Section 6.3 and is constituted of a discrete set of 6D pose obtained with the 6D pose annotation process introduced in Section 3. The performances of the methods are evaluated in Precision and Recall, using accuracy measures MPD and MSD as introduced in Section 5. More details on methods and results processing are given as supplementary material.

Results.

We evaluated SpyroPose [3] and LiePose-Diffusion [4] (either PBR + real images and PBR only to compare with SpyroPose) since they provide 6D pose distributions, unlike [9, 10] that provide distributions only over the rotation. To show genericity, we also

	Methods	Train	$\mathbf{P_{MSD}}$	$\mathbf{R_{MSD}}$	$\mathbf{P_{MPD}}$	$\mathbf{R_{MPD}}$
	T-LESS [14] **Complete dataset**					
Distrib	SpyroPose [3]	PBR	**32.8**	48.1	55.9	55.5
Distrib	LiePose-Diffusion [4]	PBR	24.6	**71.4**	**61.2**	**89.5**
Distrib	LiePose-Diffusion [4]	PBR + real	29.9	75.4	68.4	90.6
Single	leroy-fuseocclu-rgb	PBR + real	**55.4**	**35.7**	**80.9**	**54.2**
Single	gdrnppdet-pbrrealgenflow-multihypo-rgb [22, 55]	PBR + real	52.4	34.9	77.7	52.2
Single	cosypose-eccv20-syntreal-8views [20]	PBR + real	43.5	25.7	53.1	30.9
	T-LESS [14] **Only symmetrical objects**					
Distrib	SpyroPose [3]	PBR	**31.5**	42.7	70.5	60.8
Distrib	LiePose-Diffusion [4]	PBR	18.3	**61.6**	**63.6**	**87.3**
Distrib	LiePose-Diffusion [4]	PBR + real	24.1	67.0	70.3	88.1
Single	leroy-fuseocclu-rgb	PBR + real	**52.4**	**25.7**	**86.1**	**45.3**
Single	gdrnppdet-pbrrealgenflow-multihypo-rgb [22, 55]	PBR + real	45.8	20.8	76.5	37.8
Single	cosypose-eccv20-syntreal-8views [20]	PBR + real	29.1	6.5	40.0	10.5

Table 4 Comparison of pose distribution estimation methods on T-LESS using our ground truth pose distributions.

evaluated the top 3 RGB methods of single **pose estimation** task (with image-wise annotations). Their distribution is considered as a Dirac distribution on the single estimate. Results are given on the complete dataset, and the subset of objects with cylindrical symmetry (1-4, 13-18, 24 and 30).

Quantitative results are reported on Table 4 whereas qualitative results are provided in the supplementary material. It appears that LiePose-Diffusion outperforms SpyroPose, except for the Precision MSD. Lower accuracy are obtained with MSD error since those methods use RGB images only and no depth. Single pose methods outperform distribution methods in terms of Precision, as they are optimized for that. Yet, distribution methods produce much stronger Recalls, even more on only symmetrical objects, highlighting their ability to retrieve multiple meaningfull poses from the targeted image-wise distribution.

7 Limitations

Our annotation method relies on geometric analysis. Sensor resolution, sensor noise, field of view or motion blur may also affect disambiguating parts visibility. Considering these effects would improve the ground truth even further.

8 Conclusion

For simplicity, most of the 6D pose estimation benchmarks rely on a single 6D pose annotation per image, completed by a per-object symmetry pattern to transform this unique pose into a distribution. We argued that ignoring the per-image nature of the symmetry pattern is prone to bias in the resulting ground truth and performances evaluations.

We then proposed a method to annotate 6D pose distribution with a per-image analysis of the object symmetries. We illustrated that the resulting ground truth is

more accurate. Moreover, when using this ground truth to re-evaluate current state-of-the-art methods, we showed that the ranking of these methods changes drastically.

We also introduced metrics to evaluate methods that estimate a pose distribution and provided their first evaluation on real data. Such pose distribution qualification is crucial for downstream tasks. Indeed, access to multiple accurate pose solutions allows the selection of the best one for the task (e.g., considering obstacles) or, when the object is known to be asymmetrical, helps determine if the task can't be completed from the current viewpoint due to occlusion of the elements that break the symmetry and select the next best view.

Acknowledgment

This work was funded by the European Union's Horizon Europe Research and Innovation program under grant agreement nº 101226029 (CustomAI Project).

References

[1] Hodan, T., Sundermeyer, M., Labbe, Y., Nguyen, V.N., Wang, G., Brachmann, E., Drost, B., Lepetit, V., Rother, C., Matas, J.: BOP Challenge 2023 on Detection Segmentation and Pose Estimation of Seen and Unseen Rigid Objects. In: IEEE Conf. Comput. Vis. Pattern Recog. Worksh. (2024)

[2] Manhardt, F., Arroyo, D.M., Rupprecht, C., Busam, B., Birdal, T., Navab, N., Tombari, F.: Explaining the Ambiguity of Object Detection and 6D Pose from Visual Data. In: International Conference on Computer Vision (2019)

[3] Haugaard, R.L., Hagelskjær, F., Iversen, T.M.: SpyroPose: SE(3) Pyramids for Object Pose Distribution Estimation. In: International Conference on Computer Vision (2023)

[4] Hsiao, T.-C., Chen, H.-W., Yang, H.-K., Lee, C.-Y.: Confronting Ambiguity in 6D Object Pose Estimation via Score-Based Diffusion on SE(3). In: Conference on Computer Vision and Pattern Recognition (2024)

[5] Sundermeyer, M., Hodaň, T., Labbe, Y., Wang, G., Brachmann, E., Drost, B., Rother, C., Matas, J.: Bop Challenge 2022 on Detection, Segmentation and Pose Estimation of Specific Rigid Objects. In: Conference on Computer Vision and Pattern Recognition (2023)

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

[7] Bui, M., Birdal, T., Deng, H., Albarqouni, S., Guibas, L., Ilic, S., Navab, N.: 6d camera relocalization in ambiguous scenes via continuous multimodal inference. In: European Conference on Computer Vision (2020)

[8] Gilitschenski, I., Sahoo, R., Schwarting, W., Amini, A., Karaman, S., Rus, D.: Deep Orientation Uncertainty Learning Based on a Bingham Loss. In: International Conference on Learning Representations (2019)

[9] Murphy, K.A., Esteves, C., Jampani, V., Ramalingam, S., Makadia, A.: Implicit-PDF: Non-Parametric Representation of Probability Distributions on the Rotation Manifold. In: International Conference on Machine Learning (2021)

[10] Höfer, T., Kiefer, B., Messmer, M., Zell, A.: Hyperposepdf - Hypernetworks Predicting the Probability Distribution on SO(3). In: Proceedings of the IEEE (2023)

[11] Lee, J., Cho, M.: 3d equivariant pose regression via direct wigner-d harmonics prediction (2024)

[12] Yin, Y., Cai, Y., Wang, H., Chen, B.: Fishermatch: Semi-supervised rotation regression via entropy-based filtering. In: Conference on Computer Vision and Pattern Recognition (2022)

[13] Brégier, R., Devernay, F., Leyrit, L., Crowley, J.L.: Defining the Pose of Any 3D Rigid Object and an Associated Distance. International Journal of Computer Vision **126**(6) (2017)

[14] Hodaň, T., Haluza, P., Obdržálek, Š., Matas, J., Lourakis, M., Zabulis, X.: T-LESS: An RGB-D Dataset for 6D Pose Estimation of Texture-less Objects. In: IEEE Winter Conference on Applications of Computer Vision (2017)

[15] Xiang, Y., Schmidt, T., Narayanan, V., Fox, D.: PoseCNN: A Convolutional Neural Network for 6D Object Pose Estimation in Cluttered Scenes (2018)

[16] Hodaň, T., Matas, J., Obdržálek, Š.: On Evaluation of 6D Object Pose Estimation. In: European Conference on Computer Vision Workshop (2016)

[17] Drost, B., Ulrich, M., Navab, N., Ilic, S.: Model Globally, Match Locally: Efficient and Robust 3D Object Recognition. In: Conference on Computer Vision and Pattern Recognition (2010)

[18] Park, K., Patten, T., Vincze, M.: Pix2Pose: Pixel-Wise Coordinate Regression of Objects for 6D Pose Estimation. In: International Conference on Computer Vision (2019)

[19] Hodan, T., Baráth, D., Matas, J.: EPOS: Estimating 6D Pose of Objects with Symmetries. In: Conference on Computer Vision and Pattern Recognition (2020)

[20] Labbé, Y., Carpentier, J., Aubry, M., Sivic, J.: CosyPose: Consistent Multi-View Multi-Object 6D Pose Estimation. In: European Conference on Computer Vision (2020)

[21] Sundermeyer, M., Marton, Z.-C., Durner, M., Triebel, R.: Augmented Autoencoders: Implicit 3D Orientation Learning for 6D Object Detection. International Journal of Computer Vision **128** (2020)

[22] Wang, G., Manhardt, F., Tombari, F., Ji, X.: GDR-Net: Geometry-guided Direct Regression Network for Monocular 6D Object Pose Estimation. In: Conference on Computer Vision and Pattern Recognition (2021)

[23] Haugaard, R.L., Buch, A.G.: SurfEmb: Dense and Continuous Correspondence Distributions for Object Pose Estimation with Learnt Surface Embeddings. In: Conference on Computer Vision and Pattern Recognition (2022)

[24] Su, Y., Saleh, M., Fetzer, T., Rambach, J., Navab, N., Busam, B., Stricker, D., Tombari, F.: Zebrapose: Coarse to Fine Surface Encoding for 6Dof Object Pose Estimation. In: Conference on Computer Vision and Pattern Recognition (2022)

[25] Okorn, B., Xu, M., Hebert, M., Held, D.: Learning Orientation Distributions for Object Pose Estimation. In: International Conference on Intelligent Robots and Systems (2020)

[26] Klee, D., Biza, O., Platt, R., Walters, R.: Image to Sphere: Learning Equivariant Features for Efficient Pose Prediction. In: International Conference on Learning Representations (2022)

[27] Mayran de Chamisso, F., Meden, B., Tamaazousti, M.: HSPA: Hough Space Pattern Analysis as an Answer to Local Description Ambiguities for 3D Pose Estimation. In: British Machine Vision Conference (2022)

[28] Vutukur, S.R., Haugaard, R.L., Huang, J., Busam, B., Birdal, T.: Alignist: Cad-informed orientation distribution estimation by fusing shape and correspondences. In: European Conference on Computer Vision (2024)

[29] Brégier, R., Devernay, F., Leyrit, L., Crowley, J.L.: Symmetry-Aware Evaluation of 3D Object Detection and Pose Estimation in Scenes of Many Parts in Bulk. In: International Conference on Computer Vision Workshop (2017)

[30] Tyree, S., Tremblay, J., To, T., Cheng, J., Mosier, T., Smith, J., Birchfield, S.: 6-DoF Pose Estimation of Household Objects for Robotic Manipulation: An Accessible Dataset and Benchmark. In: International Conference on Intelligent Robots and Systems (2022)

[31] Drost, B., Ulrich, M., Bergmann, P., Härtinger, P., Steger, C.: Introducing MVTec ITODD - A Dataset for 3D Object Recognition in Industry. In: International Conference on Computer Vision Workshop (2017)

[32] Hodaň, T., Michel, F., Brachmann, E., Kehl, W., Buch, A.G., Kraft, D., Drost, B., Vidal, J., Ihrke, S., Zabulis, X., Sahin, C., Manhardt, F., Tombari, F., Kim,

T.-K., Matas, J., Rother, C.: BOP: Benchmark for 6D Object Pose Estimation. In: European Conference on Computer Vision Workshop (2018)

[33] Doumanoglou, A., Kouskouridas, R., Malassiotis, S., Kim, T.-K.: Recovering 6D Object Pose and Predicting Next-Best-View in the Crowd. In: Conference on Computer Vision and Pattern Recognition (2016)

[34] Hinterstoisser, S., Holzer, S., Cagniart, C., Ilic, S., Konolige, K., Navab, N., Lepetit, V.: Multimodal Templates for Real-Time Detection of Texture-Less Objects in Heavily Cluttered Scenes. In: International Conference on Computer Vision (2011)

[35] Hinterstoisser, S., Lepetit, V., Ilic, S., Holzer, S., Bradski, G., Konolige, K., Navab, N.: Model-Based Training, Detection and Pose Estimation of Texture-Less 3D Objects in Heavily Cluttered Scenes. In: Asian Conference on Computer Vision (2012)

[36] Brachmann, E., Krull, A., Michel, F., Gumhold, S., Shotton, J., Rother, C.: Learning 6D Object Pose Estimation Using 3D Object Coordinates. In: European Conference on Computer Vision (2014)

[37] Wang, P., Jung, H., Li, Y., Shen, S., Srikanth, R.P., Garattoni, L., Meier, S., Navab, N., Busam, B.: PhoCaL: A Multi-Modal Dataset for Category-Level Object Pose Estimation with Photometrically Challenging Objects. In: Conference on Computer Vision and Pattern Recognition (2022)

[38] Meden, B., Vega, P., De Chamisso, M., Fabrice, Bourgeois, S.: Introducing CEA IMSOLD - A Dataset for Multi-Scale Object Localization in Industry. In: International Conference on Robotics and Automation (2024)

[39] Kalra, A., Stoppi, G., Marin, D., Taamazyan, V., Shandilya, A., Agarwal, R., Boykov, A., Chong, T.H., Stark, M.: Towards co-evaluation of cameras hdr and algorithms for industrial-grade 6dof pose estimation. In: Conference on Computer Vision and Pattern Recognition (2024)

[40] Hodaň, T., Sundermeyer, M., Drost, B., Labbé, Y., Brachmann, E., Michel, F., Rother, C., Matas, J.: BOP Challenge 2020 on 6D Object Localization. In: European Conference on Computer Vision Workshop (2020)

[41] Iversen, T.M., Haugaard, R.L., Buch, A.G.: Ki-Pode: Keypoint-Based Implicit Pose Distribution Estimation of Rigid Objects. In: British Machine Vision Conference (2022)

[42] Periyasamy, A.S., Denninger, L., Behnke, S.: Learning Implicit Probability Distribution Functions for Symmetric Orientation Estimation from RGB Images Without Pose Labels. In: International Conference on Robotic Computing (2022)

[43] Petit, D., Bourgeois, S., Pavel, D., Gay-Bellile, V., Chabot, F., Barthe, L.: Ringnerf : Rethinking inductive biases for versatile and efficient neural fields. In: Leonardis, A., Ricci, E., Roth, S., Russakovsky, O., Sattler, T., Varol, G. (eds.) Computer Vision – ECCV 2024, pp. 148–164. Springer, Cham (2025)

[44] Johnson, J., Douze, M., Jégou, H.: Billion-scale similarity search with gpus. IEEE Transactions on Big Data (2021)

[45] Kaskman, R., Zakharov, S., Shugurov, I., Ilic, S.: HomebrewedDB: RGB-D Dataset for 6D Pose Estimation of 3D Objects. In: International Conference on Computer Vision Workshop (2019)

[46] Zhang, M., Wu, T., Wang, T., Wang, T., Liu, Z., Lin, D.: Omni6d: Large-vocabulary 3d object dataset for category-level 6d object pose estimation. In: European Conference on Computer Vision (2024)

[47] Örnek, E.P., Labbé, Y., Tekin, B., Ma, L., Keskin, C., Forster, C., Hodan, T.: Foundpose: Unseen Object Pose Estimation with Foundation Features. In: European Conference on Computer Vision (2024)

[48] Nguyen, V.N., Groueix, T., Salzmann, M., Lepetit, V.: GigaPose: Fast and Robust Novel Object Pose Estimation via One Correspondence. In: Conference on Computer Vision and Pattern Recognition (2024)

[49] Labbé, Y., Manuelli, L., Mousavian, A., Tyree, S., Birchfield, S., Tremblay, J., Carpentier, J., Aubry, M., Fox, D., Sivic, J.: MegaPose: 6D Pose Estimation of Novel Objects via Render & Compare. In: Proceedings of the 6th Conference on Robot Learning (CoRL) (2022)

[50] Moon, S., Son, H., Hur, D., Kim, S.: Genflow: Generalizable Recurrent Flow for 6D Pose Refinement of Novel Objects. In: Conference on Computer Vision and Pattern Recognition (2024)

[51] Nguyen, V.N., Groueix, T., Ponimatkin, G., Lepetit, V., Hodan, T.: CNOS: A Strong Baseline for CAD-Based Novel Object Segmentation. In: International Conference on Computer Vision Workshop (2023)

[52] Wen, B., Yang, W., Kautz, J., Birchfield, S.: Foundationpose: Unified 6D Pose Estimation and Tracking of Novel Objects. In: Conference on Computer Vision and Pattern Recognition (2024)

[53] Lin, J., Liu, L., Lu, D., Jia, K.: SAM-6D: Segment Anything Model Meets Zero-Shot 6D Object Pose Estimation. In: Conference on Computer Vision and Pattern Recognition (2024)

[54] Chen, J., Zhou, Z., Sun, M., Zhao, R., Wu, L., Bao, T., He, Z.: Zeropose: Cad-prompted zero-shot object 6d pose estimation in cluttered scenes. IEEE

Transactions on Circuits and Systems for Video Technology (2024)

[55] Liu, X., Zhang, R., Zhang, C., Fu, B., Tanq, J., Lianq, X., Tanq, J., Chenq, X., Zhanq, Y., Wang, G., Ji, X.: Gdrnpp. In: a Submission to the BOP Challenge 2022. Unpublished (2022)

[56] Zhang, R., Huang, Z., Wang, G., Liu, X., Zhang, C., Ji, X.: GPose2023. In: a Submission to the BOP Challenge 2023. Unpublished (2023)

[57] Li, Y., Mao, Y., Bala, R., Hadap, S.: Mrc-net: 6-dof pose estimation with multiscale residual correlation. In: Conference on Computer Vision and Pattern Recognition (2024)

[58] Cai, D., Heikkilä, J., Rahtu, E.: Sc6D: Symmetry-Agnostic and Correspondence-Free 6D Object Pose Estimation. In: International Conference on 3D Vision (2022)

[59] Lin, Y., Su, Y., Nathan, P., Inuganti, S., Di, Y., Sundermeyer, M., Manhardt, F., Stricker, D., Rambach, J., Zhang, Y.: Hipose: Hierarchical Binary Surface Encoding and Correspondence Pruning for RGB-D 6Dof Object Pose Estimation. In: Conference on Computer Vision and Pattern Recognition (2024)

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- BOPDistribMVAsuppMat.pdf