On the Evaluation of RGB-D-based Categorical Pose and Shape Estimation

Leonard Bruns[0000−0001−8747−6359] and Patric Jensfelt[0000−0002−1170−7162]

KTH Royal Institute of Technology,
Division of Robotics, Perception and Learning, Sweden
{leonardb,patric}@kth.se

Abstract. Recently, various methods for 6D pose and shape estimation of objects have been proposed. Typically, these methods evaluate their pose estimation in terms of average precision and reconstruction quality in terms of chamfer distance. In this work, we take a critical look at this predominant evaluation protocol, including metrics and datasets. We propose a new set of metrics, contribute new annotations for the Redwood dataset, and evaluate state-of-the-art methods in a fair comparison. We find that existing methods do not generalize well to unconstrained orientations and are actually heavily biased towards objects being upright. We provide an easy-to-use evaluation toolbox with well-defined metrics, method, and dataset interfaces, which allows evaluation and comparison with various state-of-the-art approaches (https://github.com/roym899/pose_and_shape_evaluation).

Keywords: Pose estimation · Shape reconstruction · RGB-D-based perception

1 Introduction

We consider the problem of pose and shape estimation on a per-category level. Classic grasp and motion planning methods often assume that full knowledge of pose and shape is available, making them difficult to apply with partial sensor information. Estimating the full shape and pose promises to bridge this gap from partial sensor information to an actionable representation. Although shape reconstruction itself is sufficient for some tasks, categorical pose estimation additionally provides a reference frame. This categorical reference frame could, for example, further enable alignment of objects and pose-dependent grasp computation (e.g., an upside-down mug has to be grasped differently from an upright mug).

Over the last two years, various learning-based categorical pose and shape estimation methods have been introduced. Most methods are built and evaluated on the two datasets proposed by [26]. The CAMERA dataset is a large dataset of real RGB-D tabletop scenes with synthetic objects generated on top of the table. The REAL dataset is a smaller real-world dataset of tabletop sequences with objects that have been scanned and tracked for the purpose of

evaluating categorical pose estimation. Notably, both datasets only contain upright objects, which opens the question of how well existing methods generalize to less constrained settings.

To answer this question, we contribute a set of annotations to evaluate unconstrained 6D pose and shape estimation. Our annotations consist of meshes and poses for handheld objects of three categories in the Redwood dataset [8]. In that dataset, the objects are freely rotated in front of the camera, and the orientations vary significantly more than in the datasets by [26].

Most methods evaluate pose estimation by following the same evaluation protocol as initially proposed by [26]. However, the method in [26] combines mask detection and pose estimation into a single network and therefore evaluates pose estimation with *average precision* (AP), which is a common detection metric. However, many of the subsequent methods assume the mask to be an input to their method, making AP an unnatural evaluation metric, as it makes the results unnecessarily difficult to interpret. Therefore, we propose a set of simpler metrics that use ground-truth masks and categories to evaluate pose estimation.

Pose estimation with shape reconstruction was first demonstrated by [4] and [24]. Both methods independently used the chamfer distance as their reconstruction metric, which is now commonly used by methods performing shape reconstruction. However, [23] showed that the chamfer distance is *not* a good measure of reconstruction quality and other metrics better correlate with perceived reconstruction quality. Therefore, in this work, we advocate for a new set of reconstruction metrics and propose a new evaluation protocol for both shape reconstruction and pose estimation.

To summarize, our contributions are:

- a well-defined evaluation protocol,
- a challenging set of novel annotations to evaluate unconstrained pose and shape estimation,
- a fair evaluation of various state-of-the-art methods, and
- an open source evaluation toolbox for the task of categorical pose and shape estimation.

2 Related Work

Since the introduction of the BOP benchmark suite [15], much progress has been made in the task of instance-level pose estimation, where a mesh of the target object is available. While such pose estimation with known meshes has made remarkable progress, the more general task of category-level pose estimation has only recently received more attention. Compared to instance-level pose estimation categorical pose and shape estimation is more difficult due to the large possible variations in shape and appearance even for a single category.

Wang et al. [26] introduced the first deep learning-based method to address the 6D pose estimation problem at the per-category level. They introduced two datasets: a synthetic dataset that combines real scenes with meshes from the

ShapeNet dataset [3] and a smaller real-world dataset that is mainly used for fine-tuning and evaluation. Their method is based on the *normalized object coordinate space* (NOCS), in which objects of one category have a common alignment. The projection of the NOCS coordinates in the image plane (also called NOCS map) is predicted by extending Mask R-CNN [14] with an additional head. From this prediction, the 6D pose and scale can be estimated by employing the Umeyama algorithm [25] with RANSAC [13] for outlier removal.

The NOCS map was predicted using only RGB information. Since geometry typically varies less than appearance for a fixed category, several methods were proposed to more directly incorporate the observed point set into the prediction. Chen et al. [4] introduced *canonical shape space* (CASS), which regresses the orientation and position directly from the cropped image and the set of observed points. As a byproduct of their method, they also reconstruct the full canonical point set. Tian et al. [24] introduced *shape prior deformation* (SPD), which uses a canonical point set and predicts a deformation based on the observed RGB-D information. CR-Net [27] and SGPA [5] are two extensions of the original idea of SPD. CR-Net uses a recurrent architecture to iteratively deform the canonical point set, and SGPA uses a transformer architecture to more effectively adjust the canonical point set. Recently, good results were demonstrated by only training on synthetically generated views of ShapeNet meshes [2].

Other methods such as [6] and [18] predict pose and bounding box without reconstructing the full shape of the object. For the evaluation presented in this work, we limit ourselves to methods that perform both reconstruction and pose estimation, although our evaluation protocol could in principle be used for pure pose estimation methods as well.

Aside from these RGB-D-based methods, more RGB-based methods have been proposed. Chen et al. [7] proposed an analysis-by-synthesis framework in which the latent representation of a generative model is iteratively optimized to fit the observed color image. The generative model allows to generate novel views of the object, but a full reconstruction is not readily available. Lee et al. [17] introduced a framework for estimating a mesh from an RGB image, while Engelmann et al. [10] propose reconstructing shapes in a representation-agnostic way by classifying the closest matching object from a database.

Evaluation The most established benchmark dataset for categorical pose estimation is the REAL275 dataset proposed by [26]. We will take a critical look at that dataset in Section 3.3 and show that it only evaluates a constrained set of orientations, hiding inherent difficulties of the task, such as multimodal orientation distributions. [26] also proposed *average precision* as a metric to evaluate pose estimation.

To evaluate shape reconstruction most papers currently use chamfer distance (CD) [2,4,24], which was introduced to measure the difference of point sets by [12]. However, [23] noted that CD is not robust to outliers, that is, the distance of outliers affects the metric. Therefore, the authors advocate using a robust thresholded metric such as F-score [16] to measure the quality of reconstruction.

3 Evaluation Protocol

In this section, we will discuss the existing and proposed evaluation protocol. We will start by formally defining the problem of categorical pose and shape estimation. We will then discuss metrics to evaluate and compare different solutions to this problem. Finally, we discuss the evaluation datasets.

3.1 Problem Definition

Let $\mathbf{I} \in \mathbb{R}^{H \times W \times 3}$ be an RGB image, $\mathbf{D} \in \mathbb{R}^{H \times W}$ be a depth map, and $\mathbf{P} \in \mathbb{R}^{3 \times 4}$ be the projection matrix of the associated camera. Further, let ${}^{i}\mathbf{T}_{j}$ be the homogeneous transformation matrix, that transforms a point ${}^{j}\mathbf{p}$ from frame j to frame i, that is, ${}^{i}\mathbf{p} = {}^{i}\mathbf{T}_{j}\,{}^{j}\mathbf{p}$. Note that depending on the context, ${}^{i}\mathbf{T}_{j}$ can also be interpreted as the 6D pose of frame j in frame i. Let ${}^{i}\mathbf{R}_{j}$ and ${}^{i}\mathbf{t}_{j}$ further denote the rotation matrix and translation vector of which ${}^{i}\mathbf{T}_{j}$ is composed.

We will use o to denote the object's coordinate frame and c to denote the camera's coordinate frame. We will use $\mathcal{O}$ to denote a 3D object and $\mathcal{B}(\mathcal{O})$ to denote the axis-aligned bounding box of $\mathcal{O}$ in $\mathcal{O}$'s frame o. We will assume that the center of $\mathcal{B}(\mathcal{O})$ is at the origin of frame o. We further assume that transforms can be applied to 3D objects and bounding boxes, for example, ${}^{\mathrm{c}}\mathcal{O} = {}^{\mathrm{c}}\mathbf{T}_{\mathrm{o}}\mathcal{O}$. Following this notation, note that there is a difference between $\mathcal{B}({}^{\mathrm{c}}\mathcal{O})$ and ${}^{\mathrm{c}}\mathbf{T}_{\mathrm{o}}\mathcal{B}(\mathcal{O})$. The first one is an *axis-aligned bounding box* (AABB), the second is an *oriented bounding box* (OBB).

Problem 1. (Categorical Pose and Shape Estimation) Given $(\mathbf{I}, \mathbf{D}, \mathbf{P})$ imaging an object $\mathcal{O}$ of known category c at pose ${}^{\mathrm{c}}\mathbf{T}_{\mathrm{o}}$, and given the mask $\mathbf{M}$ of visible points of the object in the image, find estimates $\tilde{\mathcal{O}}$ and ${}^{\mathrm{c}}\tilde{\mathbf{T}}_{\mathrm{o}}$ of $\mathcal{O}$ and ${}^{\mathrm{c}}\mathbf{T}_{\mathrm{o}}$, respectively.

Similarly, one could define the problems of categorical pose estimation (estimate ${}^{\mathrm{c}}\mathbf{T}_{\mathrm{o}}$ only) and categorical pose and size estimation (estimate ${}^{\mathrm{c}}\mathbf{T}_{\mathrm{o}}$ and $\mathcal{B}(\mathcal{O})$). Extensions to multiple images are possible by introducing a world coordinate system, but are not further considered in this work.

3.2 Metrics

Various metrics exist to assess how well a method solves Problem 1. Currently, the predominant evaluation metric is *average precision* [26,4,27,5,2,18,6,17] for pose estimation and *chamfer distance* [12,4,27,5,2,17] for shape reconstruction. In this section, we will introduce these metrics and discuss several issues with them. Subsequently, we will advocate for *precision* (contrary to average precision) and *F-score* to evaluate pose estimation and shape reconstruction, respectively.

We first define similarity measures for transforms and objects. These are later used to define the evaluation metrics for Problem 1.

Definition 1. *Let* $d({}^{\mathrm{c}}\mathbf{T}_{\mathrm{o}}, {}^{\mathrm{c}}\widetilde{\mathbf{T}}_{\mathrm{o}})$ *denote the translation error between the ground-truth transform and the estimated transform, that is,*

$$d({}^{\mathrm{c}}\mathbf{T}_{\mathrm{o}}, {}^{\mathrm{c}}\widetilde{\mathbf{T}}_{\mathrm{o}}) = \|{}^{\mathrm{c}}\mathbf{t}_{\mathrm{o}} - {}^{\mathrm{c}}\tilde{\mathbf{t}}_{\mathrm{o}}\|_2. \tag{1}$$

Definition 2. *Let* $\delta({}^{\mathrm{c}}\mathbf{T}_{\mathrm{o}}, {}^{\mathrm{c}}\widetilde{\mathbf{T}}_{\mathrm{o}})$ *denote the rotation error between the ground-truth transform and the estimated transform, that is,*

$$\delta({}^{\mathrm{c}}\mathbf{T}_{\mathrm{o}}, {}^{\mathrm{c}}\widetilde{\mathbf{T}}_{\mathrm{o}}) = \left| \frac{\mathrm{trace}({}^{\mathrm{c}}\mathbf{R}_{\mathrm{o}}\, {}^{\mathrm{c}}\widetilde{\mathbf{R}}_{\mathrm{o}}^{-1}) - 1}{2} \right|. \tag{2}$$

Definition 3. *Let* $\mathrm{IoU}(\mathcal{B}_1, \mathcal{B}_2)$ *denote the true intersection over union (IoU) of two tight, oriented bounding boxes [1]. Further, let the axis-aligned IoU between two objects be defined by*

$$\mathrm{IoU}^{+}({}^{\mathrm{c}}\mathcal{O}, {}^{\mathrm{c}}\widetilde{\mathcal{O}}) = \mathrm{IoU}\left(\mathcal{B}({}^{\mathrm{c}}\mathcal{O}), \mathcal{B}({}^{\mathrm{c}}\widetilde{\mathcal{O}})\right), \tag{3}$$

and the true IoU using tight, oriented bounding boxes by

$$\mathrm{IoU}({}^{\mathrm{c}}\mathcal{O}, {}^{\mathrm{c}}\widetilde{\mathcal{O}}) = \mathrm{IoU}\left({}^{\mathrm{c}}\mathbf{T}_{\mathrm{o}}\mathcal{B}(\mathcal{O}), {}^{\mathrm{c}}\widetilde{\mathbf{T}}_{\mathrm{o}}\mathcal{B}(\widetilde{\mathcal{O}})\right). \tag{4}$$

The current evaluation protocol [26] uses axis-aligned IoU^{+} instead of oriented IoU, although the former is less accurate. Our implementation follows [1] and computes oriented IoU.

Chamfer Distance In the context of shape reconstruction *chamfer distance* (CD) was introduced by [12] to differentiably measure the difference of point sets.

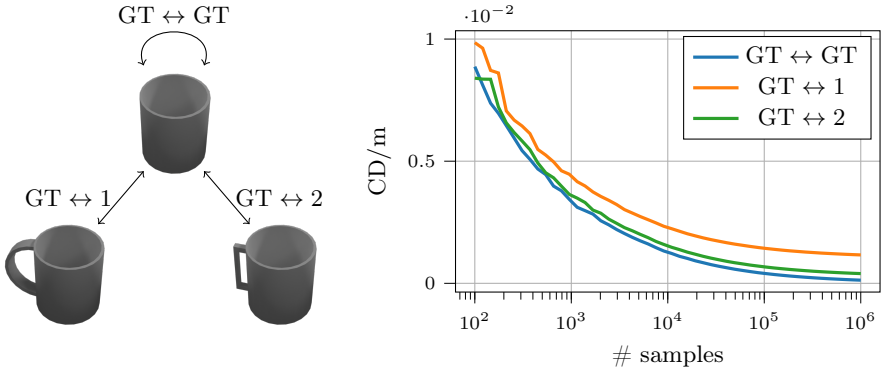
Definition 4. *Let* $\mathcal{S} \subset \mathbb{R}^3$ *and* $\widetilde{\mathcal{S}} \subset \mathbb{R}^3$ *denote point sets sampled from the surfaces of* $\mathcal{O}$ *and* $\widetilde{\mathcal{O}}$*, respectively. We define CD as*

$$\mathrm{CD}(\mathcal{S}, \widetilde{\mathcal{S}}) = \frac{1}{2|\mathcal{S}|} \sum_{\mathbf{x}\in\mathcal{S}} \min_{\mathbf{y}\in\widetilde{\mathcal{S}}} \|\mathbf{x}-\mathbf{y}\|_2 + \frac{1}{2|\widetilde{\mathcal{S}}|} \sum_{\mathbf{y}\in\widetilde{\mathcal{S}}} \min_{\mathbf{x}\in\mathcal{S}} \|\mathbf{x}-\mathbf{y}\|_2. \tag{5}$$

It is easiest to interpret as the mean Euclidean distance from a point in one point set to the closest point in the other set. Note that slightly different CD versions exist, such as squared versions and ones using the sum instead of arithmetic mean.

In Fig. 1 we visualize potential issues with CD as an evaluation metric. Consider the two mugs with different handles (denoted 1 and 2) as reconstructions of the mug without handle (denoted GT). The relative quality difference of these reconstructions measured by CD varies significantly depending on the number of samples. Furthermore, a large number of samples is required for CD to converge.

Notably, the number of ground-truth samples is left unspecified by most methods. This problem becomes further amplified because methods perform reconstruction by predicting a varying number of samples as noted by [2]. Therefore, we discourage further use of CD for evaluation purposes.



(a) Meshes (b) Chamfer distance for varying number of samples

Fig. 1: Visualization of the effect of varying number of samples on the chamfer distance. We consider two reconstructions (denoted by 1 and 2) of the ground truth (denoted by GT). Note that particularly the relative difference between $\mathrm{CD}(\mathcal{S}_{\mathrm{GT}}, \mathcal{S}_1)$ and $\mathrm{CD}(\mathcal{S}_{\mathrm{GT}}, \mathcal{S}_2)$ varies significantly. This is because the majority of the error stems from sparse sampling, not from actual differences in geometry. All mugs have been scaled to be 10 cm tall.

Reconstruction F-score Following [23], we advocate using the F-score instead of CD to evaluate shape reconstruction. Furthermore, note that we can evaluate shape reconstruction in the object frame or in the camera frame, taking into account ${}^{\mathrm{c}}\mathbf{T}_{\mathrm{o}}$ and ${}^{\mathrm{c}}\widetilde{\mathbf{T}}_{\mathrm{o}}$. Although previous methods evaluated in the object frame (i.e., assuming perfect alignment based on the canonical reference frame), we believe it is better to evaluate posed reconstruction in the camera frame, as it correlates more directly with downstream usability of the full estimate.

Definition 5. *Let $\mathcal{S} \subset \mathbb{R}^3$ and $\widetilde{\mathcal{S}} \subset \mathbb{R}^3$ denote point sets sampled from the surfaces of $\mathcal{O}$ and $\widetilde{\mathcal{O}}$, respectively. Given an application-specific threshold Δ, we define reconstruction recall as*

$$r_\Delta = \frac{1}{|\mathcal{S}|} \sum_{\mathbf{x} \in \mathcal{S}} \left[\min_{\mathbf{y} \in \widetilde{\mathcal{S}}} \|\mathbf{x} - \mathbf{y}\|_2 < \Delta \right] \tag{6}$$

and reconstruction precision as

$$p_\Delta = \frac{1}{|\widetilde{\mathcal{S}}|} \sum_{\mathbf{y} \in \widetilde{\mathcal{S}}} \left[\min_{\mathbf{x} \in \mathcal{S}} \|\mathbf{x} - \mathbf{y}\|_2 < \Delta \right], \tag{7}$$

where $[\cdot]$ denotes the Iverson bracket. Finally, we define F-score as the harmonic mean of precision and recall

$$F_\Delta = \frac{2}{p_\Delta^{-1} + r_\Delta^{-1}}. \tag{8}$$

Note that Δ should be adjusted depending on the application and the sensor. For the table-top items contained in the datasets, we propose to use $\Delta = 1\,\text{cm}$.

In Fig. 2 we show $F_{1\text{cm}}$ for varying numbers of samples for the meshes in Fig. 1(a). $F_{1\text{cm}}$ converges significantly faster than CD and can easily be interpreted as the percentage of correct (i.e., error below $\Delta = 1\,\text{cm}$) surfaces or points [23].

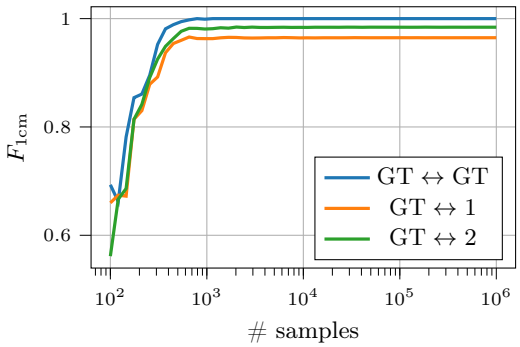


Fig. 2: Visualization of the effect of varying number of samples on the $F_{1\text{cm}}$ metric. Note that compared to CD (see Fig. 1) $F_{1\text{cm}}$ converges significantly faster.

All metrics so far (i.e., $d, \delta, \text{IoU}, \text{CD},$ and F_Δ) assess the quality of a single estimate. Next, we discuss average precision and precision, which attempt to summarize a method's performance on a dataset. In principle, one could also compute different averages of the aforementioned metrics, but those are typically affected by outliers and hard to interpret in comparison to thresholded evaluation metrics that classify estimates as true or false.

Average Precision *Average precision* (AP) summarizes precision-recall curves in a single value [22] and has been the standard evaluation metric for object detection on the PASCAL VOC [11] and COCO datasets [19]. In general, average precision is calculated based on the interpolated precision-recall curve, which is constructed by varying a confidence threshold.

Wang et al. [26] proposed to use AP with different thresholds on IoU^+, d, and δ to evaluate their pose estimation (all specified thresholds must hold for a prediction to count as true positive). Their method includes a Mask R-CNN architecture to detect objects and therefore had a confidence threshold to compute AP. However, *none* of the following pose and shape estimation methods include such a confidence threshold. Instead they all assume $\mathbf{M}$ to be given as stated in Problem 1.

To still follow the same evaluation protocol as [26], all other methods rely on the same, suboptimal Mask R-CNN predictions that [26] provided. This pro-

tocol effectively limits the achievable AP due to wrong classifications, missing detections and poor masks. Furthermore, AP is inherently difficult to interpret compared to simpler metrics.

Therefore, we believe that AP is unnecessary to compare pose and shape estimation methods. Rather, simpler metrics, such as precision (see below), should be used, assuming that mask $\mathbf{M}$ and category c are provided.

Precision We propose to use precision, contrary to average precision, to assess categorical pose and shape estimation.

Definition 6. *Given inputs* $(\mathbf{I}^i, \mathbf{D}^i, \mathbf{P}^i, \mathbf{M}^i, c^i)$*, ground truths* $(\mathcal{O}^i, {}^{\mathrm{c}}\mathbf{T}_{\mathrm{o}}^i)$*, and associated predictions* $(\widetilde{\mathcal{O}}^i, {}^{\mathrm{c}}\widetilde{\mathbf{T}}_{\mathrm{o}}^i)$ *with* $i = 1, ..., N$*, let precision be defined as*

$$P = \frac{\sum_{i=1}^{N} \left[\mathrm{c}(\mathcal{O}^i, {}^{\mathrm{c}}\mathbf{T}_{\mathrm{o}}^i, \widetilde{\mathcal{O}}^i, {}^{\mathrm{c}}\widetilde{\mathbf{T}}_{\mathrm{o}}^i) \right]}{N}, \tag{9}$$

where $[\cdot]$ *denotes the Iverson bracket and* c *determines whether a prediction is correct or not based on a single or multiple thresholds on* IoU*,* δ*,* d *or* F_Δ*.*

That is, precision measures the percentage of correct estimates based on thresholds on translation error, rotation error, IoU and F-score. Note that we can use this simpler metric instead of average precision because we decouple pose estimation from detection and classification.

Summary We propose to evaluate categorical pose and shape estimation methods by calculating precision at varying thresholds of d, δ and F_Δ. Furthermore, this evaluation procedure can be adjusted for categorical pose estimation by using d and δ only and for categorical pose and size estimation by using IoU instead of F_Δ. The thresholds for these metrics must be adjusted based on application requirements, sensor accuracy, and annotation quality. Furthermore, when combining multiple thresholds, care should be taken that they are roughly equally strict.

Note that when evaluating δ and IoU, extra care must be taken with respect to the categories containing symmetric objects. We follow [26] and ignore rotations around the up-axis for the bottle, bowl, and can categories. Further issues with ambiguities are discussed in Section 5.

3.3 Datasets

So far, most methods have been evaluated on the synthetic CAMERA25 dataset and on the smaller real-world dataset REAL275 [26]. Since we are more interested in real-world performance, we only include the REAL275 dataset in our evaluation protocol. Next, we will give an overview over the REAL275 dataset, and our new annotations for the Redwood dataset [8].

REAL275 The REAL dataset was proposed by Wang et al. [26] and consists of 4300 training images (7 video sequences) and 2750 test images (6 video sequences). The dataset contains 6 categories (bottle, bowl, camera, can, laptop, and mug) and contains 4 to 7 objects per scene. Meshes for each object are provided, obtained using an RGB-D reconstruction algorithm. Since we are primarily interested in evaluation, we will focus on the evaluation split, called REAL275, from here.

Fig. 3: Examples of REAL275 samples for the 6 object categories. Note that all objects are positioned upright on a table.

Fig. 3 shows point sets with their corresponding ground-truth annotations. Note that all objects are upright on planar surfaces. Similar constrained orientations can be found in the training splits of the CAMERA and REAL datasets. Fig. 6(a) visualizes the distribution of orientations contained in the REAL275 dataset. Note that such constraints, present in training and test data, can significantly simplify the learning problem as pose and shape ambiguities disappear (e.g., upright or upside-down can).

Redwood To evaluate methods on less constrained orientations, we contribute annotations for a set of images in the Redwood dataset [8]. The Redwood dataset contains sequences of handheld objects being freely rotated in front of the camera. No ground-truth reconstructions are provided for these sequences.

We annotated pose and shape for 3 categories (bottle, bowl, mug) for 5 sequences each. These annotations were created by manually creating OBBs in

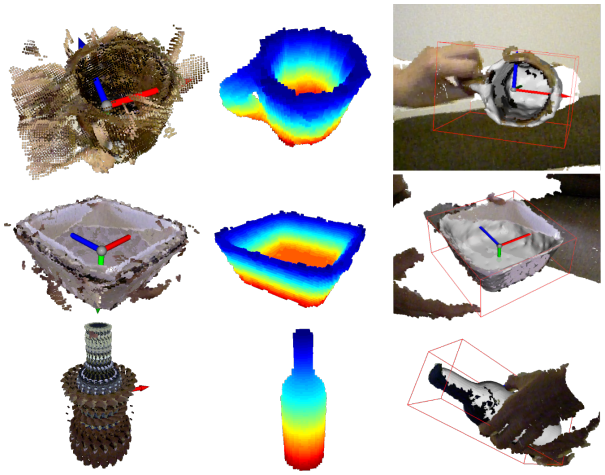


Fig. 4: Manual annotations for Redwood dataset. The left column shows the cropped, accumulated point sets (including symmetries) extracted from annotated bounding boxes. The middle columns shows the voxel grid after carving. The right column shows the extracted mesh, overlaid with the point set.

multiple frames and exploiting potential symmetries of the object. Alignment of OBBs with previous annotations was sped up and refined by using the iterative closest point (ICP) algorithm. For each of the annotated sequences we took a subset of 5 frames covering various orientations. We will refer to this set of annotations as REDWOOD75.

To reconstruct the shape from the partially occluded and noisy depth data, we start from a dense voxel grid inside the bounding box and apply voxel carving using the annotated frames to remove hands and other temporary occlusions. The remaining voxel grid contains only voxels that are not observed to be free in any of the annotated frames. From this voxel grid, we extract a mesh and apply Laplacian smoothing. Fig. 4 visualizes the annotation process.

Note that this method only approximates the real shape and is sensitive to misaligned bounding boxes and missing depth data. Especially thin surfaces and details, such as mug handles, are difficult to extract accurately due to sensor noise. Additionally, alignment errors can easily accumulate, resulting in too large or too small objects. However, the annotations are accurate enough to evaluate the performance of current methods on unconstrained orientations.

To produce the final ground truth, we compute tight bounding boxes based on the extracted meshes. Fig. 5 shows examples of the final annotations. In Fig. 6 we compare the orientation distribution of REDWOOD75 and REAL275. Note that the orientations in REDWOOD75 are significantly less constrained than in REAL275.

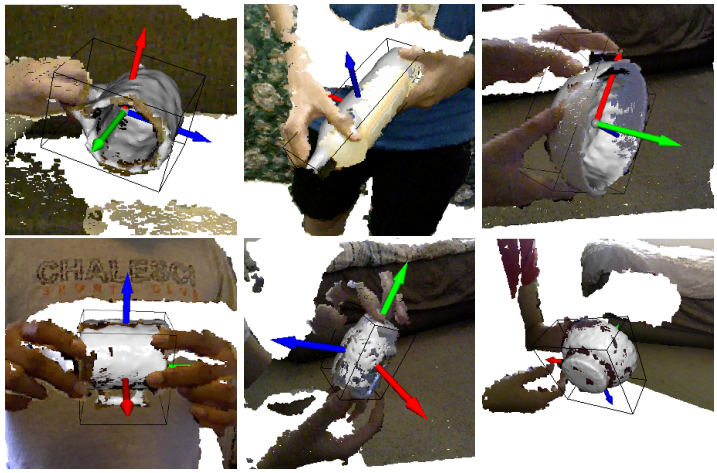
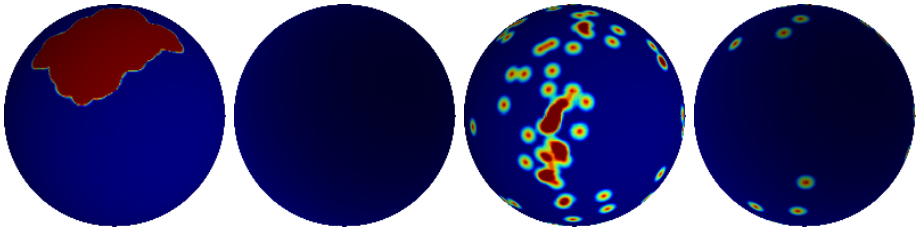


Fig. 5: Examples of REDWOOD75 samples.



(a) REAL275 (b) REDWOOD75

Fig. 6: Distribution of the up-axis in REAL275 and REDWOOD75 datasets. REDWOOD75 covers a larger variety of orientations.

4 Experiments

We follow our proposed evaluation protocol and compare three methods for shape and pose estimation: CASS [4], SPD [24], and ASM-Net [2]. CASS and SPD are well-established baselines for this task and were both trained on the CAMERA and REAL datasets. Unlike the other methods, ASM-Net [2] was trained on synthetic ShapeNet [3] renderings only. All methods estimate 6D pose and reconstruct a point set of varying density.

For all methods, we closely followed the published inference code and verified that our method interface produced results similar to their evaluation code. We found that only SPD's published model achieved the same qualitative results as shown in their publication. CASS' reconstructed point sets were significantly worse except for the laptop category and ASM-Net often predicted negative scales, which causes some reconstructions to be upside down, while the object frame o is predicted in the correct orientation.

We have implemented the metrics, and interfaces to the datasets and methods described in the previous sections using Open3D [28] and PyTorch [21]. We open-source our code as a benchmarking toolbox, with the goal of simplifying fair comparison with state-of-the-art methods. We plan to extend the toolbox as new methods are released.

Qualitative Results Fig. 7 shows randomly selected results on the REAL275 and REDWOOD75 datasets. On REAL275, all methods perform pose estimation with a similar quality as shown in the respective publications. On REDWOOD75, on the other hand, only ASM-Net shows limited generalization capability. CASS and SPD predict upright objects consistent with the orientation distribution of REAL275 independent of the input.

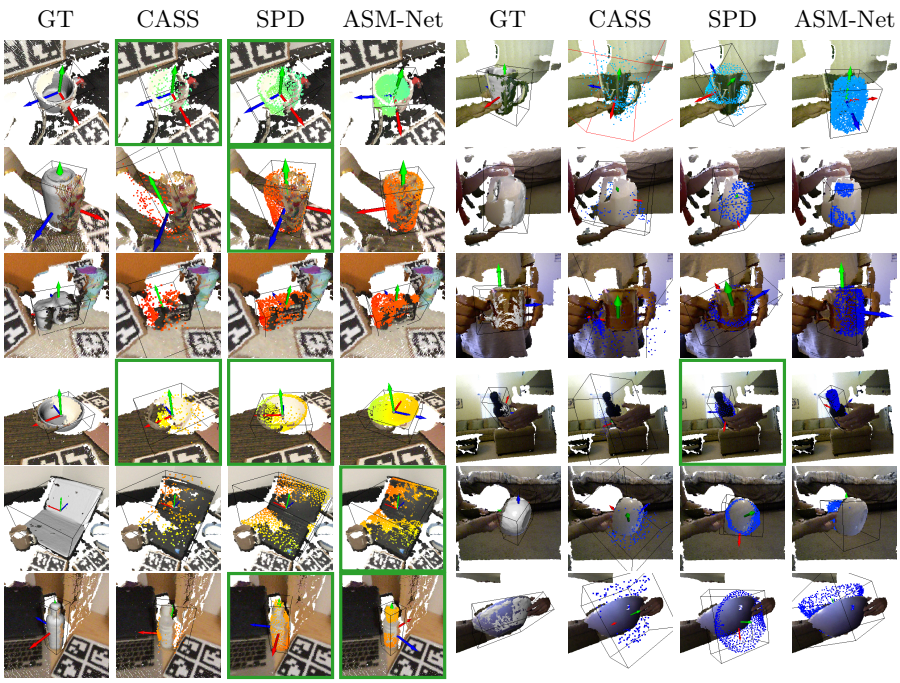
The shape reconstructions of SPD are qualitatively the best. ASM-Net's reconstructions are often flipped, but surfaces typically align well. As noted above, CASS completely fails to reconstruct any object except laptops. For both ASM-Net and CASS it is unclear whether this performance difference stems from errors in the code or if the published model weights are suboptimal.

Table 1: Precision at varying position, orientation and F-score thresholds.

	REAL275			REDWOOD75		
	CASS	SPD	ASM-Net	CASS	SPD	ASM-Net
10°,2 cm	0.331	**0.535**	0.331	0.013	0.2	**0.307**
5°,1 cm	0.073	**0.205**	0.069	0.000	0.013	**0.080**
10°,2 cm,0.6	0.031	**0.471**	0.215	0.000	**0.173**	**0.173**
5°,1 cm,0.8	0.000	**0.170**	0.050	0.000	0.013	**0.053**

Quantitative Results We now present the results using the metrics introduced in Section 3.2. In Table 1 we report precision with thresholds of varying strictness. To assess pose estimation independent of shape estimation, we use $5°, 1\,\text{cm}$ and $10°, 2\,\text{cm}$. To further include shape reconstruction, we use $5°, 1\,\text{cm}, 0.8$ and $10°, 2\,\text{cm}, 0.6$ for δ, d, and $F_{1\text{cm}}$, respectively. We picked these tuples of thresholds such that all thresholds in a tuple are roughly equally strict. In the past, some methods used pairs such as $10°, 5\text{cm}$, or $10°, 10\text{cm}$, where only the $10°$ threshold practically mattered.

The results from Table 1 confirm the qualitative observations from before. CASS in particular performs poorly on the shape reconstruction metrics. Note that for both datasets there is still a lot of room for improvement. Typically, significantly fewer than 50% of the estimates are of sufficient quality to be considered correct in pose and shape. This shows that categorical pose and shape estimation is still an open problem, especially for unconstrained orientations. The performance gap between the two datasets is especially noticeable for SPD, which performs better than ASM-Net on REAL275 but worse on REDWOOD75.



(a) REAL275 (b) REDWOOD75

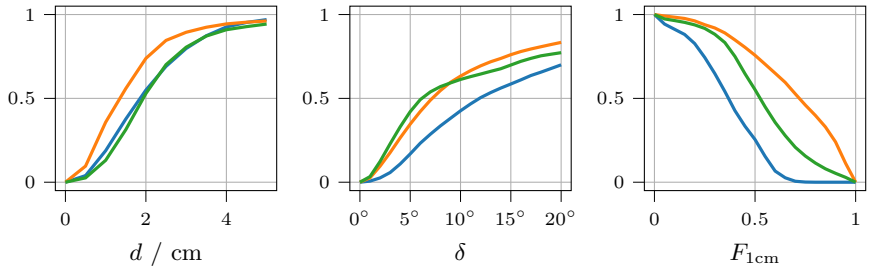
Fig. 7: Randomly selected results on REAL275 and REDWOOD75 datasets. Results that are considered correct under $\delta = 10°, d = 2\,\text{cm}, F_{1\text{cm}} = 0.6$ thresholds are highlighted.

To gain further insight into the estimation quality of the methods, we show detailed results for varying thresholds in Fig. 8. It can be seen that the difference between the two datasets is more pronounced for orientation-based thresholding. This confirms the issue of constrained orientations discussed in Section 3.3 (see Fig. 6). ASM-Net, which was trained on synthetic data (the exact distribution was not specified but is likely less constrained than the CAMERA and REAL datasets), performs best in this metric.

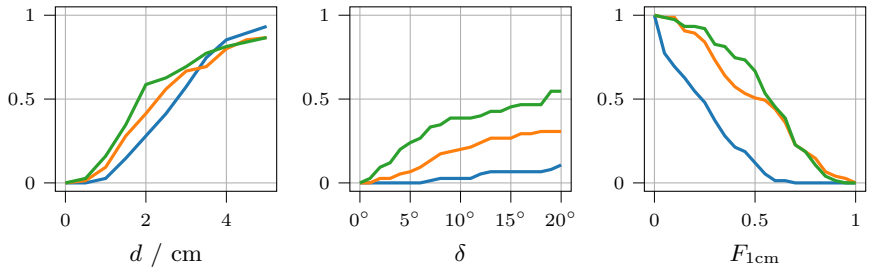
In Table 2 we further report the precision per category with the more lenient pose and shape estimation thresholds $\delta = 10°, d = 2\,\text{cm}, F_{1\text{cm}} = 0.6$. Note that CASS' reconstructions are very sparse and noisy, and therefore rarely reach $F_{1\text{cm}} > 0.6$ (see also Fig. 8). On REAL275, all methods fail at the camera category which contains significantly more shape variation than the other categories. Note that, despite these relatively lenient thresholds, all methods fail to sufficiently recover pose and shape for most of the REDWOOD75 samples.

Run Time We also compute the mean run time of each method per prediction. We include the transfer time from CPU to GPU, but exclude the computation

of the metrics. In our experiments[1], CASS required 22.3 ms, SPD 199 ms, and ASM-Net 58.6 ms.



(a) REAL275



(b) REDWOOD75

Fig. 8: Detailed precision results for CASS (—), SPD (—) and ASM-Net (—) for varying thresholds of position, orientation and F-score thresholds.

Table 2: Precision for different categories with $\delta = 10°, d = 2\,\text{cm}, F_{1\text{cm}} = 0.6$.

	REAL275						REDWOOD75		
	Bottle	Bowl	Camera	Can	Laptop	Mug	Bottle	Bowl	Mug
CASS	0.002	0.093	0.001	0.030	0.0	0.068	0.000	0.000	0.000
SPD	**0.610**	**0.892**	**0.052**	**0.863**	**0.218**	**0.246**	**0.320**	0.160	**0.040**
ASM-Net	0.167	0.137	0.023	0.587	0.133	0.229	0.160	**0.360**	0.000

5 Limitations

Comparability The results from Section 4 suggest that training on synthetic data currently generalizes better to unconstrained orientations. This is expected,

[1] Intel Core i7-6850K CPU, NVIDIA TITAN X (Pascal) GPU

since synthetic data allows accurate, unconstrained generation of training data. This opens the question of whether CASS or SPD would generalize better or worse when trained on the same synthetic data as ASM-Net. Since the training code for CASS and ASM-Net has not been published, it is difficult to perform further comparisons. In general, since methods currently vary significant parts of training datasets, architecture, pose parameterization, and losses, it is difficult to assess the individual contributions of a single change.

Multimodal Distributions Unconstrained pose estimation introduces significant difficulties in the task, which were hidden due to the constraints present in the CAMERA and REAL datasets. Consider, for example, the bottom left mug in Fig. 5. From the given view, it is difficult to tell which way the opening of the mug faces. Another example are cans, which are geometry-wise nearly symmetric. Currently there are only few works [20,9] that consider this problem of ambiguous poses and evaluation of such methods that predict multimodal posteriors is difficult. One possible way of evaluating such methods would be to allow methods to generate N hypotheses. Precision could then be computed for the best and worst hypothesis. A strong method would generate the same hypothesis N times if there is no ambiguity. If there is ambiguity, the correct hypothesis would still be contained in the set of hypotheses with a high probability. Alternatively, likelihood-based metrics might be suitable if methods provide likelihoods instead of hypotheses.

Dataset Size The REDWOOD75 dataset is limited in size, but the results suggest a clear lack of generalization capability of current approaches. This shows the need for larger datasets for unconstrained pose and shape estimation. It is an open question how such a dataset could be collected in the most efficient way.

6 Conclusion and Outlook

In this work, we have discussed the limitations of the current evaluation protocol prevalent in the field of categorical pose and shape estimation. In particular, existing datasets contain only a heavily constrained set of orientations, which simplifies the problem by removing pose and shape ambiguities. Furthermore, existing evaluation metrics are suboptimal and unnecessarily difficult to interpret. To alleviate these problems, we propose a new set of metrics applicable to both the established REAL275 and our proposed REDWOOD75 dataset, which contains a large variety of orientations. We apply our evaluation protocol to three methods and confirm limited generalization capability as suggested by the constrained orientations in their training data.

Our experiments suggest that there is a need for larger, high-quality datasets for unconstrained pose and shape estimation as well as for methods that can handle unconstrained orientations and the resulting pose ambiguities in a principled way.

Acknowledgements This work was partially supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation.

References

1. Ahmadyan, A., Zhang, L., Ablavatski, A., Wei, J., Grundmann, M.: Objectron: A large scale dataset of object-centric videos in the wild with pose annotations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7822–7831 (2021)
2. Akizuki, S., Hashimoto, M.: Asm-net: Category-level pose and shape estimation using parametric deformation. In: Proceedings of the British Machine Vision Conference (2021)
3. Chang, A.X., Funkhouser, T., Guibas, L., Hanrahan, P., Huang, Q., Li, Z., Savarese, S., Savva, M., Song, S., Su, H., Xiao, J., Yi, L., Yu, F.: ShapeNet: An information-rich 3D model repository. Tech. Rep. 1512.03012, arXiv preprint (Dec 2015)
4. Chen, D., Li, J., Wang, Z., Xu, K.: Learning canonical shape space for category-level 6d object pose and size estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11973–11982 (2020)
5. Chen, K., Dou, Q.: Sgpa: Structure-guided prior adaptation for category-level 6d object pose estimation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2773–2782 (2021)
6. Chen, W., Jia, X., Chang, H.J., Duan, J., Shen, L., Leonardis, A.: Fs-net: Fast shape-based network for category-level 6d object pose estimation with decoupled rotation mechanism. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1581–1590 (2021)
7. Chen, X., Dong, Z., Song, J., Geiger, A., Hilliges, O.: Category level object pose estimation via neural analysis-by-synthesis. In: European Conference on Computer Vision. pp. 139–156. Springer (2020)
8. Choi, S., Zhou, Q.Y., Miller, S., Koltun, V.: A large dataset of object scans. arXiv preprint arXiv:1602.02481 (2016)
9. Deng, H., Bui, M., Navab, N., Guibas, L., Ilic, S., Birdal, T.: Deep bingham networks: Dealing with uncertainty and ambiguity in pose estimation. arXiv preprint arXiv:2012.11002 (2020)
10. Engelmann, F., Rematas, K., Leibe, B., Ferrari, V.: From points to multi-object 3d reconstruction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4588–4597 (2021)
11. Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The PASCAL visual object classes (VOC) challenge. International Journal of Computer Vision 88(2), 303–338 (2010)
12. Fan, H., Su, H., Guibas, L.J.: A point set generation network for 3d object reconstruction from a single image. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 605–613 (2017)
13. Fischler, M.A., Bolles, R.C.: 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)
14. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask R-CNN. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2961–2969 (2017)

15. 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. pp. 577–594. Springer (2020)
16. Knapitsch, A., Park, J., Zhou, Q.Y., Koltun, V.: Tanks and temples: Benchmarking large-scale scene reconstruction. ACM Transactions on Graphics 36(4) (2017)
17. Lee, T., Lee, B.U., Kim, M., Kweon, I.S.: Category-level metric scale object shape and pose estimation. IEEE Robotics and Automation Letters 6(4), 8575–8582 (2021)
18. Lin, J., Wei, Z., Li, Z., Xu, S., Jia, K., Li, Y.: Dualposenet: Category-level 6d object pose and size estimation using dual pose network with refined learning of pose consistency. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3560–3569 (2021)
19. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: Common objects in context. In: Proceedings of the European Conference on Computer Vision. pp. 740–755 (2014)
20. 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: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6841–6850 (2019)
21. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: PyTorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems 32, 8026–8037 (2019)
22. Salton, G., McGill, M.J.: Introduction to modern information retrieval. McGraw Hill (1983)
23. Tatarchenko, M., Richter, S.R., Ranftl, R., Li, Z., Koltun, V., Brox, T.: What do single-view 3d reconstruction networks learn? In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3405–3414 (2019)
24. Tian, M., Ang, M.H., Lee, G.H.: Shape prior deformation for categorical 6d object pose and size estimation. In: European Conference on Computer Vision. pp. 530–546. Springer (2020)
25. Umeyama, S.: Least-squares estimation of transformation parameters between two point patterns. IEEE Transactions on Pattern Analysis and Machine Intelligence 13(04), 376–380 (1991)
26. Wang, H., Sridhar, S., Huang, J., Valentin, J., Song, S., Guibas, L.J.: Normalized object coordinate space for category-level 6d object pose and size estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2642–2651 (2019)
27. Wang, J., Chen, K., Dou, Q.: Category-level 6d object pose estimation via cascaded relation and recurrent reconstruction networks. In: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems. pp. 4807–4814 (2021)
28. Zhou, Q.Y., Park, J., Koltun, V.: Open3d: A modern library for 3d data processing. arXiv preprint arXiv:1801.09847 (2018)