MegaPose: 6D pose estimation of novel objects via render&compare

Anonymous Author(s)
Affiliation
Address
email

Abstract: We introduce MegaPose, a method to estimate the 6D pose of novel objects, i.e., objects unseen during training, in an image. The method only assumes knowledge of (i) a region of interest displaying the object in the image and (ii) a CAD model of the observed object at *inference time*. The contributions of this work are threefold. First, we present a 6D pose refiner based on a render & compare strategy which can be applied to novel objects. The shape and coordinate system of the novel object are provided as inputs to the network by rendering multiple synthetic views of the object's CAD model. Second, we introduce a novel approach for coarse pose estimation which leverages a network trained to classify whether the pose error between a synthetic rendering and an observed image of the same object can be corrected by the refiner. Third, we introduce a large scale synthetic dataset of photorealistic images of thousands of objects with diverse visual and shape properties, and show that this diversity is crucial to obtain good generalization performance on novel objects. We train our approach on this large synthetic dataset and apply it *without retraining* to hundreds of novel objects in real images from several pose estimation benchmarks. Our approach achieves state-of-the-art performance on the ModelNet dataset. An extensive evaluation on the 7 core datasets of the BOP challenge demonstrates that our approach achieves performance competitive with existing approaches that require access to the target objects during training.

Keywords: Object Pose Estimation, Robot Manipulation, Computer Vision

1 Introduction

Accurate 6D object pose estimation is essential for many robotic and augmented reality applications. Current state-of-the-art methods are learning-based [1, 2, 3, 4, 5] and require 3D models of the objects of interest at both training and test time. These methods require hours (or days) to generate synthetic data for each object and train the pose estimation model. They thus cannot be used in the context of robotic applications where the objects are only known during inference (e.g. CAD models are provided by a manufacturer or reconstructed [6]), and where rapid deployment to novel scenes and objects is key.

The goal of this work is to estimate the 6D pose of novel objects, *i.e.*, objects that are only available at *inference time* and are not known in advance during training. This problem presents the challenge of generalizing to the large variability in shape, texture, lighting conditions, and severe occlusions that can be encountered in real-world applications. Some prior works [7, 8, 9, 10, 11, 12, 13] have considered category-level pose estimation to partially address the challenge of novel objects by developing methods that can generalize to novel object instances of a known class (e.g. mugs or shoes). These methods however do not generalize to object instances outside of training categories. Other methods aim at generalizing to any novel instances regardless of their category [14, 15, 16, 17, 18, 19, 20, 21]. These works present important technical limitations. They rely on non-learning based components for generating pose hypotheses [20] (e.g. PPF [22]), for pose refinement [16] (e.g. PnP [23] and ICP [24, 25]), for computing photometric errors in pixel space [14], or for estimating the object depth [17, 15] (e.g. using only the size of a 2D detection [26]). These components depend on several hyper-parameters, and inherently cannot benefit from being trained on large amount of data to gain robustness with respect to noise, occlusions, or object variability.

Submitted to the 6th Conference on Robot Learning (CoRL 2022). Do not distribute.

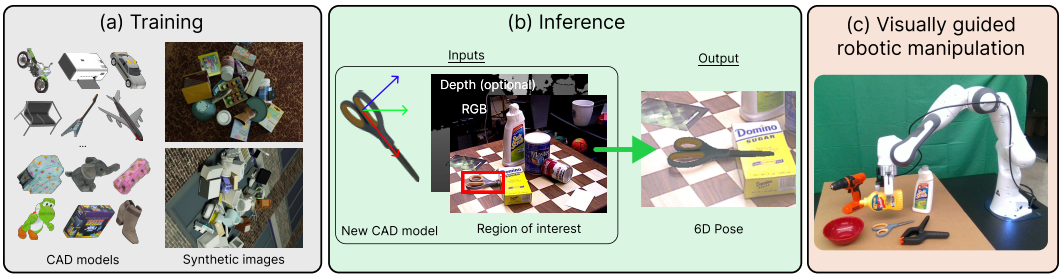

Figure 1: MegaPose is a 6D pose estimation approach (a) that is trained on millions of synthetic scenes with thousands of different objects and (b) can be applied *without re-training* to estimate the pose of any novel object, given a CAD model and a region of interest displaying the object. It can thus be used to rapidly deploy visually guided robotic manipulation systems in novel scenes containing novel objects (c).

Pipelines for 6D pose estimation of known (not novel) objects that consist of multiple learned stages [3, 4] have shown excellent performance on several benchmarks [1] with various illumination conditions, textureless objects, cluttered scenes and high levels of occlusions. We take inspiration from [3, 4] which split the problem into three parts: (i) 2D object detection, (ii) coarse pose estimation, and (iii) iterative refinement via render & compare. We aim at extending this approach to novel objects unseen at training time. The detection of novel objects has been addressed by prior works [16, 27, 28, 29] and is outside the scope of this paper. In this work, we focus on the coarse and refinement networks for 6D pose estimation. Extending the paradigm from [3] presents three major challenges. First, the pose of an object depends heavily on both its visual appearance and choice of coordinate system (defined in the CAD model of the object). In existing refinement networks based on render & compare [19, 3], this information is encoded in the network weights during training, leading to poor generalization results when tested on novel objects. Second, direct regression methods for coarse pose estimation are trained with specific losses for symmetric objects [3], requiring that object symmetries be known in advance. Finally, the diversity of shape and visual properties of the objects that can be encountered in real-world applications is immense. Generalizing to novel objects requires robustness to properties such as object symmetries, variability of object shape, and object textures (or absence of).

Contributions. We address these challenge and propose a complete method for estimating the pose of any novel object in a single RGB or RGB-D image, as illustrated in Figure 1. First, we propose a novel approach for 6D pose refinement based on render & compare which enables generalization to novel objects. The shape and coordinate system of the novel object are provided as inputs to the network by rendering multiple synthetic views of the objects CAD model. Second, we propose a novel method for coarse pose estimation which does not require knowledge of the object symmetries during training. The coarse pose estimation is formulated as a classification problem where we compare renderings of random pose hypotheses with an observed image, and decide whether the pose can be corrected by the refiner. Finally, we leverage the availability of large-scale 3D model datasets to generate a highly diverse synthetic dataset consisting of 2 million photorealistic [30] images depicting over 20K models in physically plausible configurations. The code, dataset and pre-trained models will be released upon publication.

We show that our novel-object pose estimation methods trained on our large-scale synthetic dataset achieves state-of-the-art performance on ModelNet [31, 19]. We also perform an extensive evaluation of the approach on hundreds of novel objects from all 7 core datasets of the BOP challenge [1] and demonstrate that our approach achieves performance competitive with existing approaches that require access to the target objects during training.

2 Related work

In this section, we first review the literature on 6D pose estimation of known rigid objects. We then focus on the practical scenario similar to ours where the objects are not known in advance for training.

6D pose estimation of known objects. Estimating the 6D pose of rigid objects is an old problem [32, 33, 34] that was first addressed using correspondences established with locally invari-

ant features [34, 35, 36, 37, 22] or template matching [38, 39]. These have been replaced by learning-based methods based on convolutional neural networks (CNNs) that directly regress sets of sparse [40, 41, 26, 42, 43, 44, 45] or dense [46, 47, 48, 49, 2, 43] features. All these approaches use non-learning stages relying on PnP+Ransac [50, 23] to recover the pose from correspondences in RGB images, or variants of the iterative closest point algorithm (ICP) [24, 25] when depth is available. The best performing methods rely on trainable refinement networks [51, 19, 3, 19, 4] based on render & compare [52, 53, 54, 19]. These methods render a single image of the object, which is not sufficient to provide complete information on the shape and coordinate system of a 3D model to the network. This information is thus encoded in the networks weights when training, which leads to poor generalization when tested on novel objects unseen at training. Our approach renders multiple views of an object to provide this 3D information, making the trained network independent of these object-specific properties.

6D pose estimation of novel objects. Other works consider a practical scenario where the objects are not known in advance. Category-level 6D pose estimation is a popular problem [7, 8, 9, 10, 11, 12, 13] in which CAD models of test objects are not known, but the objects are assumed to belong to a known category. These methods rely on object properties that are common within categories (e.g. handle of a mug) to define and estimate the object pose, and thus cannot generalize to novel categories. Our method requires the 3D model of the novel object instance to be known during inference, but does not rely on any category-level information. Other works address a scenario similar to ours. [55, 18, 56, 17, 15, 29] only estimate the 3D orientation of novel objects by comparing rendered pose hypotheses with the observed image using features extracted by a network. They rely on handcrafted [17, 15] or learning-based DeepIM [18] refiners to recover accurate 6D poses. We instead propose a method that estimates the full 6D pose of the object and show our refinement network significantly outperforms DeepIM [19] when tested on novel object instances. The closest work to ours are OSOP [16] and Zephyr [20]. OSOP focuses on the coarse estimation by explicitly predicting 2D-2D correspondences between a single rendered view of the object and the observed image, and solves for the pose using PnP or Kabsch [24] which makes inference slower and less robust compared to directly predicting refinement transforms with a network as done in our solution. Zephyr [20] strongly relies on the depth modality, whereas our approach can also be used in RGB-only images. Finally, [14, 57, 21, 58] investigate using a set of real reference views of the novel object instead of using a CAD model. These approaches have only reported results on simple datasets with limited or no occlusions. Our use of a deep render & compare network trained on a large-scale synthetic dataset displaying highly occluded object instances enables us to handle highly cluttered scenes with high occlusions like in the LineMOD Occlusion, HomebrewedDB or T-LESS datasets.

3 Method

In this section we present our framework for pose estimation of novel objects. Our goal is to detect the pose $\mathcal{T}_{CO}$ (the pose of object frame O expressed in camera frame C composed of 3D rotation and 3D translation) of a novel object given an input RGB (or RGBD) image, I_o, and a 3D model of the object. Similar to DeepIM [19] and CosyPose [3], our method consists of three components (1) object detection, (2) coarse pose estimation and (3) pose refinement. Compared to these works, our proposed method enables generalization to novel objects not seen during training, requiring novel approaches for the coarse model, the refiner and the training data. Our approach can accept either RGB or RGBD inputs. If depth is available the RGB and D images are concatenated before being passed into the network. Detection of novel-objects in an image is an interesting problem that has been addressed in prior work [27, 59, 16, 21, 29] but lies outside the scope of this paper. Thus for our experiments we assume access to an object detector, but emphasize that our method can be coupled with any object detector, including zero-shot methods such as those in [27, 59].

3.1 Technical Approach

Coarse pose estimation. Given an object detection, shown in Figure 1(b), the goal of the coarse pose estimator is to provide an initial pose estimate $\mathcal{T}_{CO,\text{coarse}}$ which is sufficiently accurate that it can then be further improved by the refiner. In order to generalize to novel-objects we propose a novel

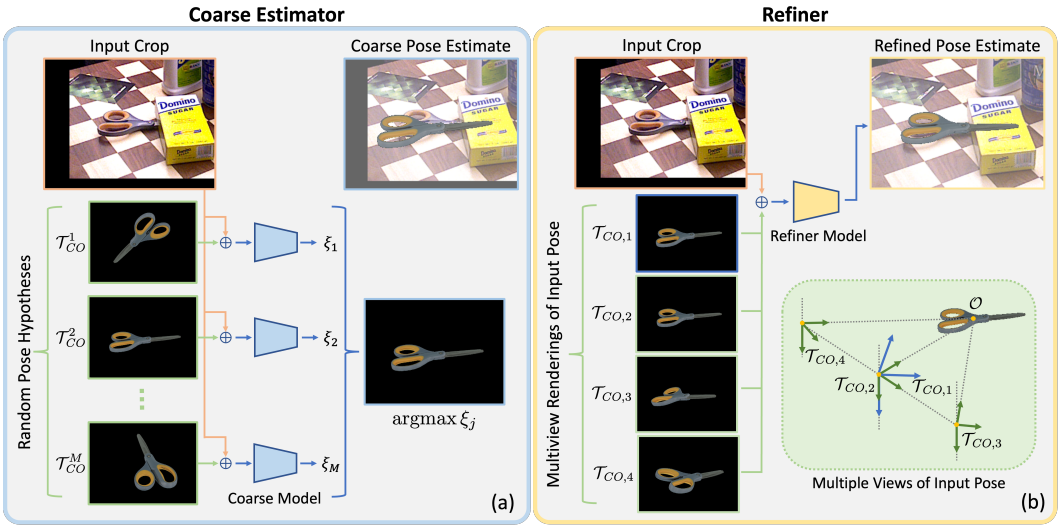

Figure 2: $\oplus$ denotes concatenation. **(a) Coarse Estimator:** Given a cropped input image the coarse module renders the object in multiple input poses $\{\mathcal{T}_{CO}^j\}$. The coarse network then classifies which rendered image best matches the observed image. **(b) Refiner:** Given an initial pose estimate $\mathcal{T}_{CO}^k$ the refiner renders the objects at the estimated pose $\mathcal{T}_{CO,1} := \mathcal{T}_{CO}^k$ along with 3 additional viewpoints $\{\mathcal{T}_{CO,i}\}_{i=2}^4$ defined such that the camera z-axis intersects the anchor point $\mathcal{O}$. The refiner network consumes the concatenation of the observed and rendered images and predicts an updated pose estimate $\mathcal{T}_{CO}^{k+1}$.

classification based approach that compares observed and rendered images of the object in a variety of poses and selects the rendered image whose object pose best matches the observed object pose.

Figure 2(a) gives an overview of the coarse model. At inference time the network consumes the observed image I_o along with rendered images $\{I_r(\mathcal{T}_{CO}^j)\}_{j=1}^M$ of the object in many different poses $\{\mathcal{T}_{CO}^j\}_{k=1}^M$. For each pose $\mathcal{T}_{CO}^j$ the model predicts a score $(I_o, I_r(\mathcal{T}_{CO}^j)) \rightarrow \xi_j$ that classifies whether the pose hypothesis is within the basin of attraction of the refiner. The highest scoring pose $\mathcal{T}_{CO}^{j^*}, j* = \text{argmax}_j \xi_j$ is used as the initial pose for the refinement step. Since we are performing classification, our method can implicitly handle object symmetries, as multiple poses can be classified as correct.

Pose refinement model. Given an input image and an estimated pose, the refiner predicts an updated pose estimate. Starting from a coarse initial pose estimate $\mathcal{T}_{CO,\text{coarse}}$ we can iteratively apply the refiner to produce an improved pose estimate. Similar to [3, 19] our refiner takes as input observed I_o and rendered images $I_r(\mathcal{T}_{CO}^k)$ and predicts an updated pose estimate $\mathcal{T}_{CO}^{k+1}$, see Figure 2 (b), where k refers to the kth iteration of the refiner. Our pose update uses the same parameterization as DeepIM [19] and CosyPose [3] which disentangles rotation and translation prediction. Crucially this pose update $\Delta\mathcal{T}$ depends on the choice of an *anchor point* $\mathcal{O}$, see Appendix for more details . In prior work [3, 19] which trains and tests on the same set of objects, the network can effectively learn the position of the anchor point $\mathcal{O}$ for each object. However in order to generalize to novel objects we must enable the network to infer the anchor point $\mathcal{O}$ at inference time.

In order to provide information about the anchor point to the network we always render images $I_r(\mathcal{T}_{CO}^k)$ such that the anchor point $\mathcal{O}$ projects to the image center. Using rendered images from multiple distinct viewpoints $\{\mathcal{T}_{CO,i}\}_{i=1}^N$ the network can infer the location of the anchor point $\mathcal{O}$ as the intersection point of camera rays that pass through the image center, see Figure 2(b).

Additional information about object shape and geometry can be provided to the network by rendering depth and surface normal channels in the rendered image I_r. We normalize both input depth (if available) and rendered depth images using the currently estimated pose $\mathcal{T}_{CO}^k$ to assist the network in generalizing across object scales, see Appendix for more details.

Network architecture. Both the coarse and refiner networks consists of a ResNet-34 backbone followed by spatial average pooling. The coarse model has a single fully-connected layer that consumes

the backbone feature and outputs a classification logit. The refiner network has a single fully-connected layer that consumes the backbone feature and outputs 9 values that specify the translation and rotation for the pose update.

3.2 Training Procedure

Training data. For training, both the coarse and refiner models require RGB(-D)[1] images with ground-truth 6D object pose annotations, along with 3D models for these objects. In order for our approach to generalize to novel-objects we require a large dataset containing diverse objects. All of of our methods are trained purely on synthetic data generated using BlenderProc [30]. We generate a dataset of 2 million images using a combination of ShapeNet [60] (abbreviated as SN) and Google-Scanned-Objects (abbreviated as GSO) [6]. Similar to the BOP [61] synthetic data, we randomly sampled objects from our dataset and dropped them on a plane using a physics simulator. Materials, background textures, lighting and camera positions are randomized. Example images can be seen in Figure 1(a) and in the Appendix. Some of our ablations also use the synthetic training datasets provided by the BOP challenge [61]. We add data augmentation similar to CosyPose [3] to the RGB images which was shown to be a key to successful sim-to-real transfer. We also apply data augmentation to the depth images as explained in the appendix.

Refiner model. The refiner model is trained similarly to [3]. Given an image with an object $\mathcal{M}$ at ground-truth pose $\mathcal{T}^*_{CO}$ we generate a perturbed pose $\mathcal{T}'_{CO}$ by applying a random translation and rotation to $\mathcal{T}^*_{CO}$. Translation is sampled from a normal distribution with a standard deviations of $(0.02, 0.02, 0.05)$ centimeters and rotation is sampled as random Euler angles with a standard deviation of 15 degrees in each axis. The network is trained to predict the relative transformation between the initial and target pose. Following [3, 19] we use a loss that disentangles the prediction of depth, x-y translation, and rotation. See the appendix for more details.

Coarse model. Given an input image I_o of an object $\mathcal{M}$ and a pose $\mathcal{T}'_{CO}$ the coarse model is trained to classify whether pose $\mathcal{T}'_{CO}$ is within the basin of attraction of the refiner. In other words, if the refiner were started with the initial pose estimate $\mathcal{T}'_{CO}$ would it be able to estimate the ground-truth pose via iterative refinement? Given a ground-truth pose-annotation $\mathcal{T}^*_{CO}$ we randomly sample poses $\mathcal{T}'_{CO}$ by adding random translation and rotation to $\mathcal{T}^*_{CO}$. The positives are sampled from the same distribution used to generate the perturbed poses the refiner network is trained to correct (see above), and other poses sufficiently distinct to this one (see the appendix for more details) are marked as negatives. The model is then trained with binary cross entropy loss.

4 Experiments

We evaluate our method for 6D pose estimation of novel objects using the seven challenging datasets of the BOP [1, 61] 6D pose estimation benchmark, and the ModelNet [19] dataset. The dataset and the standard 6D pose estimation metrics we use are detailed in Section 4.1. In all our experiments, the objects are considered novel, i.e. they are only available during inference on a new image and they are not used during training. In Section 4.2, we evaluate the performance of our approach composed of coarse and refinement networks. Notably, we show that (i) our method is competitive with others that require the object models to be known in advance, and (ii) our refiner outperforms current state-of-the-art on the ModelNet dataset. Section 4.3 validates our technical contributions and shows the crucial importance of the training data in the success of our method. Finally, we discuss the limitations in Section 4.4.

4.1 Dataset and metrics

We consider the seven core datasets of the BOP challenge [61, 1]: LineMod Occlusion (LM-O) [62], T-LESS [63], TUD-L [61], IC-BIN [64], ITODD [65], HomebrewedDB (HB) [66] and YCB-Video (YCB-V) [46]. These datasets exhibit 132 different objects in cluttered scenes with occlusions. These objects present many factors of variation: textured or untextured, symmetric or

[1] Our method can consume either RGB or RGB-D images depending on the input modalities that are available.

	Pose Initialization		Pose Refinement			BOP Datasets							
	Method	Novel objects	Method	Novel objects	RGB-D Input	LM-O	T-LESS	TUD-L	IC-BIN	ITODD	HB	YCB-V	Mean
1	CosyPose	✗	CosyPose	✗		63.3	64.0	68.5	58.3	21.6	65.6	57.4	57.0
2	SurfEmb	✗	BFGS	✗		66.3	73.5	71.5	58.8	41.3	79.1	64.7	65.0
3	SurfEmb	✗	BFGS+ICP	✓	✓	*75.8*	*82.8*	*85.4*	*65.6*	*49.8*	*86.7*	*80.6*	*75.2*
4	OSOP	✓	Multi-Hyp.	✓		31.2	-	-	-	-	49.2	33.2	-
5	OSOP	✓	MH+ICP	✓	✓	48.2	-	-	-	-	60.5	57.2	-
6	Drost,PPF	✓	MH+ICP	✓	✓	52.7	-	-	-	-	67.1	34.4	-
7	Drost,PPF	✓	Zephyr+ICP	✓	✓	**59.8**	-	-	-	-	-	51.6	-
8	CosyPose	✗	–			53.6	52.0	57.6	53.0	13.1	33.5	33.3	42.3
9	CosyPose	✗	Ours	✓		65.5	72.0	70.1	57.3	28.4	67.0	56.8	59.6
10	CosyPose	✗	Ours	✓	✓	71.2	63.8	85.0	55.1	39.9	73.2	69.2	66.0
11	Ours	✓	–			18.7	19.7	20.5	15.3	8.00	18.6	13.9	16.2
12	Ours	✓	Ours	✓		53.7	**62.2**	58.4	**43.6**	30.1	72.9	60.4	54.5
13	Ours	✓	Ours	✓	✓	58.3	54.3	**71.2**	37.1	**40.4**	**75.7**	**63.3**	**57.2**

Table 1: Results on the BOP challenge datasets. We report the AR score on each of the 7 datasets considered in the BOP challenge and the mean score across datasets. With the exception of Zephyr (row 11), all approaches are trained purely on synthetic data. For each column, we denote the best over result in *italics* and the best novel-object pose estimation method in **bold**.

		Average Recall		
Method	RGB-D	(5°, 5cm)	ADD (0.1d)	Proj2D (5px)
DeepIM [19]	✓	64.3	83.6	73.3
Multi-Path [15]		84.8	90.1	81.6
LatentFusion[14]	✓	85.5	94.3	94.7
Ours		88.6	90.5	88.9
Ours	✓	**97.6**	**98.9**	**97.5**

Table 2: Evaluation of the refiner on the ModelNet [19] dataset. The mean average recall is computed over the seven classes of the dataset.

asymmetric, household or industrial (e.g. watcher pitcher, stapler, bowls, multi-socket plug adaptor) which makes them representative of objects that are typically encountered in robotic scenarios. The ModelNet dataset depicts individual instances of objects from 7 classes of the ModelNet [31] dataset (bathtub, bookshelf, guitar, range hood, sofa, tv stand and wardrobe). Initial object poses are provided and the focus is on refining these inital poses. We follow the evaluation protocol of [1] for BOP datasets, and of DeepIM [19] for ModelNet.

4.2 6D pose estimation of novel objects

Performance of coarse+refiner. Table 1 reports results of our novel-object pose estimation method on the BOP datasets. Note that since novel instance detection is outside the scope of this paper, we use the detections from Mask-RCNN networks trained on the synthetic PBR data [1] publicly available for each dataset. This makes it difficult to directly compare with other approaches for pose estimation of novel objects [20, 16] where the focus is on the detector and coarse estimates unlike our work which focuses on predicting fine 6D pose estimates of novel objects using learning-based coarse and refinement networks. We report the results of our coarse model (l.11) (i.e. line 11 of Table 1), and after running 5 iterations of the refiner network, on RGB (l.12) or RGB-D (l.13) images. We observe that (i) our refinement network significantly improves the coarse estimates (+41 avg. AR score for our RGBD refiner) and (ii) the performance of both models is competitive with the learning-based refiner of CosyPose [3] (l.1) while not requiring to be trained on the test objects. The recent SurfEmb [2] performs better than our approach, but heavily relies on the knowledge of the objects for training and cannot generalize to novel ones.

Performance of the refiner. We now focus on the evaluation of our refiner which can be used to refine arbitrary initial poses. Our refiner is the only learning-based approach in Table 1 which can be applied to novel objects. In lines 9 and 10, we apply our refiner to the coarse estimates of CosyPose [3] (l.11). Again, we observe that our refiner significantly improves the accuracy of these initial pose estimates

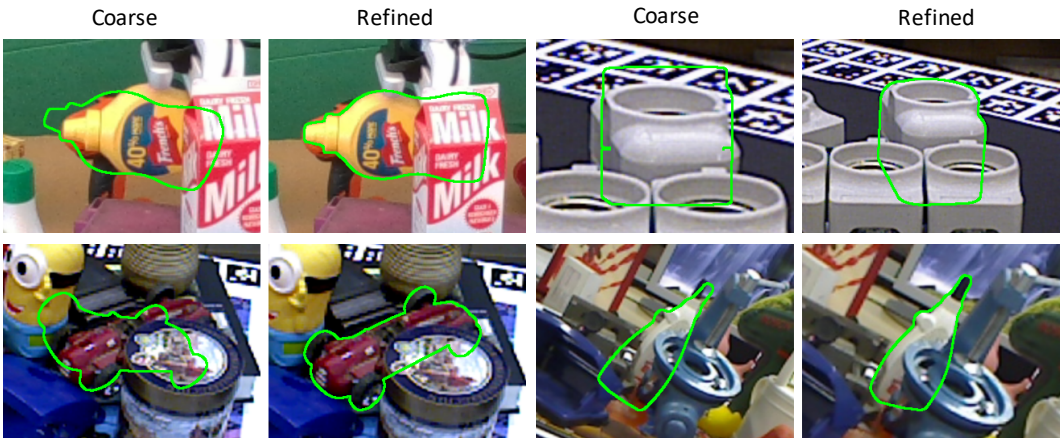

Figure 3: Qualitative results. For each pair of images, the left image is a visualization of our coarse estimate, and the right image is after applying 5 iterations of our refiner. None of these objects from the YCBV, LMO, HB, or T-LESS datasets were used for training our approach. Please notice the high accuracy of MegaPose despite (i) severe occlusions and (ii) the varying properties of the novel objects (e.g. the texture-less industrial plug in the top-right example, textured mustard bottle in the top-left.).

Rendered views	Rendered normals	BOP5	ModelNet ADD(0.1d)
1	✓	52.0	83.3
2	✓	59.0	90.4
4	✓	**61.7**	**96.1**
4		59.1	83.1

(a)

Training objects	Num. objects	BOP5	ModelNet ADD(0.1d)
GSO+ShapeNet	10 + 100	47.9	28.7
GSO+ShapeNet	100 + 1000	49.3	80.3
GSO+ShapeNet	250 + 2000	56.9	82.0
GSO+ShapeNet	500 + 10000	59.3	89.3
GSO+ShapeNet	1000 + 20000	61.7	**96.1**
GSO	1000	62.2	95.7
BOP	132	**62.6**	93.4

(b)

Table 3: Ablation study. We study (a) using multiple rendered object views and normal maps as input to our RGB-only refiner model and (b) training the refiner on different variations of our large-scale synthetic dataset. Average recall is reported on BOP5 (mean of LM-O, T-LESS, TUD-L, IC-BIN and YCB-V) and ModelNet.

(+23.7 in avg for the RGB-D model). Notably, the RGB-only method (l.9) performs better than the CosyPose refiner (l.1) in average, while not having seen the BOP objects during training. This is thanks to our large-scale training on thousands of various objects, while CosyPose is only trained on tens of objects for each dataset. One iteration of our refiner takes approximately 50 milliseconds on a RTX 2080 GPU, making it suitable for use in an online tracking application. Five iterations of our refiner is also 5 times faster than the object-specific refiner of SurfEmb [2] which takes around 1 second per image crop. Finally, we evaluate our refiner on ModelNet and compare it with the state-of-the-art methods MP-AAE [15] and LatentFusion [14]. For this experiment, we remove the ShapeNet categories that overlap with the test ones in ModelNet from our training set in order to provide a fair comparison on novel instances and novel categories similar to [14, 15, 19]. Results reported in Table 2 show that our refiner significantly outperforms existing approaches across all metrics.

4.3 Ablations

In this section we perform ablations of our approach to validate our main contributions. Additional ablations are in the appendix. For these ablations, we consider the RGB-only refiner and re-train several models with different configurations of hyper-parameters and training data.

Encoding the anchor point and object shape. As discussed in Section 3.1 the refiner must have information about the anchor point $\mathcal{O}$ in order to generalize to novel objects. We accomplish this by using 4 rendered views pointing towards the anchor point, see Figure 2(b). Table 3(a) shows the

performance of the refiner increases as we increase the number of views from 1 to 4, validating our design choice. Multiple views may also help the network to understand the object's appearance from alternate viewpoints, thus potentially helping the refiner to overcome large initial pose errors. We also validate our choice to provide a normal map of the object to the network. This information can help the network use subtle object appearance variations that are only visible under different illumination like the details on the cross of a guitar.

Classification-based coarse network. To validate our classification-based coarse model, we consider a regression-based alternative. We trained a regression-based network similar to the coarse model of CosyPose [3] which takes as input six views of the objects covering viewpoints at the poles of a sphere centered on the object. The network collapsed during training, leading to large errors that cannot be recovered by the refiner and a performance close to zero on the BOP datasets. We hypothesize this failure is due to the presence of symmetric objects in our training set which leads to ambiguous gradients during training.

Number of training objects. We now show the crucial role of the training data. We report in Table 3 (b) the results for our refiner trained on an increasing number of CAD models. The performance steadily increases with the number of objects, which validates that training on a large number of object models is important to generalize to novel ones. These results also suggests that the performance of our approach *could* be improved as more datasets of high-quality CAD models like GSO [6] become available.

Variety in the training objects. Next, we restrict the training to different sets of objects. We observe in the bottom of Table 3(b) that models from the GoogleScannedObjects are more important to the performance of the method on the BOP dataset compared to using both ShapeNet and GSO. We hypothesize this is due to the presence of high-quality textured objects in the GSO dataset. Finally, we train our model on the 132 objects of the BOP dataset. When testing on the same BOP objects, the performance benefits of knowing these objects during training is small compared to using our GSO+ShapeNet or GSO.

4.4 Limitations

While MegaPose shows promising results in robot experiments (**please see the supplementary video**) and 6D pose estimation benchmarks, there is still room for improvement. The most common failure mode is due to inaccurate initial pose estimates from the coarse model. The refiner model is trained to correct poses that are within a constrained range, but can fail if the initial error is too large. There exist multiple potential approaches to alleviate this problem. We can increase the number of pose hypotheses M at the expense of increased inference time, improve the accuracy of the coarse model, and increase the basin of attraction for refinement model. Another limitation is the runtime of our coarse model. We use $M = 520$ pose hypotheses per object which takes around 4 seconds to be rendered and evaluated by our coarse model. In a tracking scenario however, the coarse model is run just once at the initial frame and the object can be tracked using the refiner which runs at 20Hz. Additionally, our refiner could also be coupled with alternate coarse estimation approaches such as [16, 17] to achieve improved runtime performance.

5 Conclusion

We propose MegaPose, a method for 6D pose estimation of novel objects. Megapose can estimate the 6D pose of novel objects given a CAD model of the object available only at test time. We quantitatively evaluated MegaPose on hundreds of different objects depicted in cluttered scenes, and performed ablation studies to validate our network design choices and highlight the importance of the training data. We will release our models and large-scale synthetic dataset to stimulate the development of novel methods that are practical to use in the context of robotic manipulation where rapid deployment to new scenes with new objects is crucial. While this work focuses on the coarse estimation and fine refinement of an object pose, detecting any unknown object given only a CAD model is still a difficult problem that remains to be solved for having a complete framework for detection and pose estimation of novel objects. Future work will address zero-shot object detection using our large-scale synthetic dataset.

References

[1] T. Hodaň, M. Sundermeyer, B. Drost, Y. Labbé, E. Brachmann, F. Michel, C. Rother, and J. Matas. BOP Challenge 2020 on 6D object localization. In *ECCVW*, 2020.

[2] R. L. Haugaard and A. G. Buch. Surfemb: Dense and continuous correspondence distributions for object pose estimation with learnt surface embeddings. In *CVPR*, 2022.

[3] Y. Labbé, J. Carpentier, M. Aubry, and J. Sivic. CosyPose: Consistent multi-view multi-object 6D pose estimation. In *ECCV*, 2020.

[4] L. Lipson, Z. Teed, A. Goyal, and J. Deng. Coupled iterative refinement for 6D multi-object pose estimation. In *CVPR*, 2022.

[5] G. Wang, F. Manhardt, F. Tombari, and X. Ji. GDR-Net: Geometry-guided direct regression network for monocular 6D object pose estimation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16611–16621, June 2021.

[6] L. Downs, A. Francis, N. Koenig, B. Kinman, R. Hickman, K. Reymann, T. B. McHugh, and V. Vanhoucke. Google scanned objects: A high-quality dataset of 3D scanned household items. In *ICRA*, 2022.

[7] Y. Lin, J. Tremblay, S. Tyree, P. A. Vela, and S. Birchfield. Single-stage keypoint-based category-level object pose estimation from an RGB image. In *ICRA*, 2022.

[8] H. Wang, S. Sridhar, J. Huang, J. Valentin, S. Song, and L. J. Guibas. 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*, pages 2642–2651, 2019.

[9] X. Chen, Z. Dong, J. Song, A. Geiger, and O. Hilliges. Category level object pose estimation via neural analysis-by-synthesis. In *European Conference on Computer Vision*, pages 139–156. Springer, 2020.

[10] F. Li, I. Shugurov, B. Busam, M. Li, S. Yang, and S. Ilic. Polarmesh: A star-convex 3d shape approximation for object pose estimation. *IEEE Robotics and Automation Letters*, 7(2):4416–4423, 2022.

[11] F. Manhardt, G. Wang, B. Busam, M. Nickel, S. Meier, L. Minciullo, X. Ji, and N. Navab. CPS++: Improving class-level 6D pose and shape estimation from monocular images with self-supervised learning. *arXiv preprint arXiv:2003.05848*, 2020.

[12] L. Manuelli, W. Gao, P. Florence, and R. Tedrake. KPAM: Keypoint affordances for category-level robotic manipulation. In *ISRR*, 2019.

[13] B. Wen, W. Lian, K. Bekris, and S. Schaal. Catgrasp: Learning category-level task-relevant grasping in clutter from simulation. *arXiv preprint arXiv:2109.09163*, 2021.

[14] K. Park, A. Mousavian, Y. Xiang, and D. Fox. LatentFusion: End-to-end differentiable reconstruction and rendering for unseen object pose estimation. In *CVPR*, 2020.

[15] M. Sundermeyer, M. Durner, E. Y. Puang, Z.-C. Marton, N. Vaskevicius, K. O. Arras, and R. Triebel. Multi-path learning for object pose estimation across domains. In *CVPR*, June 2020.

[16] I. Shugurov, F. Li, B. Busam, and S. Ilic. OSOP: A multi-stage one shot object pose estimation framework. In *CVPR*, 2022.

[17] V. N. Nguyen, Y. Hu, Y. Xiao, M. Salzmann, and V. Lepetit. Templates for 3D object pose estimation revisited: Generalization to new objects and robustness to occlusions. In *CVPR*, 2022.

[18] Y. Xiao, X. Qiu, P.-A. Langlois, M. Aubry, and R. Marlet. Pose from shape: Deep pose estimation for arbitrary 3D objects. In *BMVC*, 2019.

[19] Y. Li, G. Wang, X. Ji, Y. Xiang, and D. Fox. DeepIM: Deep iterative matching for 6D pose estimation. In *ECCV*, 2018.

[20] B. Okorn, Q. Gu, M. Hebert, and D. Held. ZePHyR: Zero-shot pose hypothesis rating. In *ICRA*, 2021.

[21] Y. Liu, Y. Wen, S. Peng, C. Lin, X. Long, T. Komura, and W. Wang. Gen6D: Generalizable model-free 6-DoF object pose estimation from RGB images. In *arXiv:2204.10776*, 2022.

[22] B. Drost, M. Ulrich, N. Navab, and S. Ilic. Model globally, match locally: Efficient and robust 3d object recognition. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 998–1005. Ieee, 2010.

[23] V. Lepetit, F. Moreno-Noguer, and P. Fua. EPnP: An accurate O(n) solution to the PnP problem. *IJCV*, 2009.

[24] P. J. Besl and N. D. McKay. Method for registration of 3-d shapes. In *Sensor fusion IV: control paradigms and data structures*, volume 1611, pages 586–606. Spie, 1992.

[25] Z. Zhang. Iterative point matching for registration of free-form curves and surfaces. *International journal of computer vision*, 13(2):119–152, 1994.

[26] W. Kehl, F. Manhardt, F. Tombari, S. Ilic, and N. Navab. SSD-6D: Making RGB-based 3D detection and 6D pose estimation great again. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1521–1529, 2017.

[27] A. Osokin, D. Sumin, and V. Lomakin. Os2d: One-stage one-shot object detection by matching anchor features. In *European Conference on Computer Vision*, pages 635–652. Springer, 2020.

[28] J.-P. Mercier, M. Garon, P. Giguere, and J.-F. Lalonde. Deep template-based object instance detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1507–1516, 2021.

[29] Y. Xiao and R. Marlet. Few-shot object detection and viewpoint estimation for objects in the wild. In *European Conference on Computer Vision (ECCV)*, 2020.

[30] M. Denninger, M. Sundermeyer, D. Winkelbauer, Y. Zidan, D. Olefir, M. Elbadrawy, A. Lodhi, and H. Katam. Blenderproc. *arXiv preprint arXiv:1911.01911*, 2019.

[31] Z. Wu, S. Song, A. Khosla, F. Yu, L. Zhang, X. Tang, and J. Xiao. 3d shapenets: A deep representation for volumetric shapes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1912–1920, 2015.

[32] L. G. Roberts. *Machine perception of three-dimensional solids*. PhD thesis, Massachusetts Institute of Technology, 1963.

[33] D. G. Lowe. Three-dimensional object recognition from single two-dimensional images. *Artif. Intell.*, 1987.

[34] D. G. Lowe. Object recognition from local scale-invariant features. In *ICCV*, 1999.

[35] H. Bay, T. Tuytelaars, and L. Van Gool. SURF: Speeded up robust features. In *ECCV*, 2006.

[36] A. Collet and S. S. Srinivasa. Efficient multi-view object recognition and full pose estimation. In *ICRA*, 2010.

[37] A. Collet, M. Martinez, and S. S. Srinivasa. The MOPED framework: Object recognition and pose estimation for manipulation. *The International Journal of Robotics Research*, 2011.

[38] S. Hinterstoisser, S. Holzer, C. Cagniart, S. Ilic, K. Konolige, N. Navab, and V. Lepetit. Multimodal templates for real-time detection of texture-less objects in heavily cluttered scenes. In *ICCV*, 2011.

[39] F. Jurie and M. Dhome. A simple and efficient template matching algorithm. In *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001*, volume 2, pages 544–549. IEEE, 2001.

[40] M. Rad and V. Lepetit. BB8: A scalable, accurate, robust to partial occlusion method for predicting the 3D poses of challenging objects without using depth. In *CVPR*, 2017.

[41] J. Tremblay, T. To, B. Sundaralingam, Y. Xiang, D. Fox, and S. Birchfield. Deep object pose estimation for semantic robotic grasping of household objects. In *Conference on Robot Learning (CoRL)*, 2018.

[42] B. Tekin, S. N. Sinha, and P. Fua. Real-time seamless single shot 6d object pose prediction. In *CVPR*, 2018.

[43] S. Peng, Y. Liu, Q. Huang, X. Zhou, and H. Bao. Pvnet: Pixel-wise voting network for 6dof pose estimation. In *CVPR*, 2019.

[44] G. Pavlakos, X. Zhou, A. Chan, K. G. Derpanis, and K. Daniilidis. 6-dof object pose from semantic keypoints. In *ICRA*, 2017.

[45] Y. Hu, J. Hugonot, P. Fua, and M. Salzmann. Segmentation-driven 6D object pose estimation. In *CVPR*, 2019.

[46] Y. Xiang, T. Schmidt, V. Narayanan, and D. Fox. PoseCNN: A convolutional neural network for 6D object pose estimation in cluttered scenes. In *RSS*, 2018.

[47] K. Park, T. Patten, and M. Vincze. Pix2Pose: Pixel-wise coordinate regression of objects for 6D pose estimation. In *ICCV*, 2019.

[48] C. Song, J. Song, and Q. Huang. Hybridpose: 6D object pose estimation under hybrid representations. In *CVPR*, 2020.

[49] S. Zakharov, I. Shugurov, and S. Ilic. Dpod: 6d pose object detector and refiner. In *CVPR*, 2019.

[50] R. I. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, ISBN: 0521540518, second edition, 2004.

[51] C. Wang, D. Xu, Y. Zhu, R. Martín-Martín, C. Lu, L. Fei-Fei, and S. Savarese. Densefusion: 6d object pose estimation by iterative dense fusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3343–3352, 2019.

[52] K. Pauwels and D. Kragic. Simtrack: A simulation-based framework for scalable real-time object pose detection and tracking. In *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1300–1307. IEEE, 2015.

[53] M. Oberweger, P. Wohlhart, and V. Lepetit. Training a feedback loop for hand pose estimation. In *ICCV*, 2015.

[54] F. Manhardt, W. Kehl, N. Navab, and F. Tombari. Deep model-based 6d pose refinement in rgb. In *ECCV*, 2018.

[55] M. Aubry, D. Maturana, A. A. Efros, B. Russell, and J. Sivic. Seeing 3D chairs: Exemplar part-based 2D-3D alignment using a large dataset of CAD models. In *CVPR*, 2014.

[56] Y. Xiao, Y. Du, and R. Marlet. PoseContrast: Class-agnostic object viewpoint estimation in the wild with pose-aware contrastive learning. In *3DV*, 2021.

[57] Y. He, Y. Wang, H. Fan, J. Sun, and Q. Chen. FS6D: Few-shot 6D pose estimation of novel objects. In *CVPR*, 2022.

[58] J. Sun, Z. Wang, S. Zhang, X. He, H. Zhao, G. Zhang, and X. Zhou. OnePose: One-shot object pose estimation without CAD models. In *CVPR*, 2022.

[59] E. Olson. AprilTag: A robust and flexible visual fiducial system. In *ICRA*, 2011.

[60] A. X. Chang, T. Funkhouser, L. Guibas, P. Hanrahan, Q. Huang, Z. Li, S. Savarese, M. Savva, S. Song, H. Su, J. Xiao, L. Yi, and F. Yu. ShapeNet: An Information-Rich 3D Model Repository. Technical Report arXiv:1512.03012 [cs.GR], Stanford University — Princeton University — Toyota Technological Institute at Chicago, 2015.

[61] T. Hodan, F. Michel, E. Brachmann, W. Kehl, A. GlentBuch, D. Kraft, B. Drost, J. Vidal, S. Ihrke, X. Zabulis, et al. Bop: Benchmark for 6d object pose estimation. In *ECCV*, 2018.

[62] S. Hinterstoisser, V. Lepetit, S. Ilic, S. Holzer, G. Bradski, K. Konolige, and N. Navab. Model based training, detection and pose estimation of texture-less 3d objects in heavily cluttered scenes. In *Asian conference on computer vision (ACCV)*, 2012.

[63] T. Hodan, P. Haluza, Š. Obdržálek, J. Matas, M. Lourakis, and X. Zabulis. T-LESS: An RGB-D dataset for 6D pose estimation of Texture-Less objects. In *Winter Conference on Applications of Computer Vision (WACV)*, 2017.

[64] A. Doumanoglou, R. Kouskouridas, S. Malassiotis, and T.-K. Kim. Recovering 6d object pose and predicting next-best-view in the crowd. In *CVPR*, 2016.

[65] B. Drost, M. Ulrich, P. Bergmann, P. Hartinger, and C. Steger. Introducing mvtec itodd-a dataset for 3d object recognition in industry. In *ICCV Workshops*, 2017.

[66] R. Kaskman, S. Zakharov, I. Shugurov, and S. Ilic. Homebreweddb: Rgb-d dataset for 6d pose estimation of 3d objects. In *ICCV Workshops*, 2019.