Occlusion-Robust Object Pose Estimation with Holistic Representation

Bo Chen
The University of Adelaide
bo.chen@adelaide.edu.au

Tat-Jun Chin
The University of Adelaide
tat-jun.chin@adelaide.edu.au

Marius Klimavicius
Blackswan Technologies
marius@blackswan.ltd

Abstract

Practical object pose estimation demands robustness against occlusions to the target object. State-of-the-art (SOTA) object pose estimators take a two-stage approach, where the first stage predicts 2D landmarks using a deep network and the second stage solves for 6DOF pose from 2D-3D correspondences. Albeit widely adopted, such two-stage approaches could suffer from novel occlusions when generalising and weak landmark coherence due to disrupted features. To address these issues, we develop a novel occlude-and-blackout batch augmentation technique to learn occlusion-robust deep features, and a multi-precision supervision architecture to encourage holistic pose representation learning for accurate and coherent landmark predictions. We perform careful ablation tests to verify the impact of our innovations and compare our method to SOTA pose estimators. Without the need of any post-processing or refinement, our method exhibits superior performance on the LINEMOD dataset. On the YCB-Video dataset our method outperforms all non-refinement methods in terms of the ADD(-S) metric. We also demonstrate the high data-efficiency of our method. Our code is available at http://github.com/BoChenYS/ROPE

1. Introduction

Object pose estimation is the task of inferring the relative orientation and position between the target object and the observer. Such inference is crucial in many vision applications such as robotic manipulation [67, 66, 11], augmented reality [39, 12], autonomous driving [10, 61, 63] and spacecraft navigation [6, 49]. The problem can be simplified if depth information is available [40, 58, 17, 9]. However, depth sensors are not always practical. Pose estimation from images is thus an important research problem.

In this paper we consider the problem of object pose estimation from a single RGB image. Our focus lies in the base estimator, *i.e.*, from input image to the output pose, before any refinement step. For the base estimator, a number of works [27, 62, 45, 13] adopt direct regression approaches
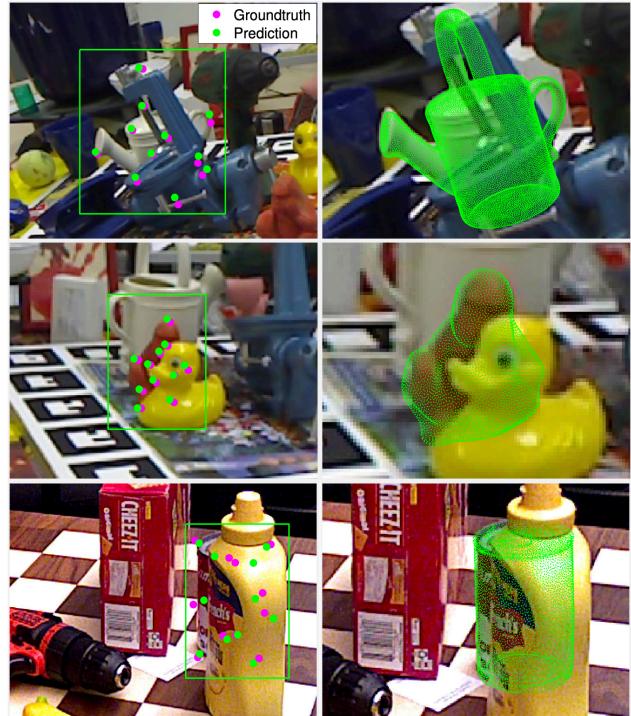

Figure 1. Qualitative results of our robust object pose estimation on the Occluded-LINEMOD (top and middle) and the YCB-Video (bottom) datasets. **Left**: prediction of bounding boxes and landmarks of the target object in a test image (zoomed view). **Right**: prediction of 6DOF poses without post-processing or refinement.

which map the input image that contains the target object to its 6 DOF pose. However, such approaches tend to be sensitive to occlusions and are observed to be similar to performing image retrieval [48].

Rather than directly regressing the pose, two-stage approaches [23, 25, 35, 41, 42, 44, 46, 64, 51, 43, 57] first predict landmarks on the object to establish 2D-3D correspondences, then use a Perspective-n-Point (PnP) like algorithm to solve for the pose. Previous results suggest that two-stage methods are generally more accurate [41, 22]. Their strengths derive from training the model with richer supervision signals (*i.e.*, groundtruth landmarks) rather than just

the pose, and injecting tolerance towards inaccurate landmark predictions by robust PnP.

However, two-stage approaches are not intrinsically immune to occlusion. Current works to improve robustness often take the pixel-wise or patch-wise approach [44, 41, 23, 25, 35, 42], *i.e.*, generating an ensemble of predictions from each image pixel or patch, and aggregate them to obtain a more robust final prediction. Although ensembling can mitigate some occlusion-induced inaccuracies, landmark coherence are easily disrupted by large and novel occlusions, because the network predicts landmarks independently and consistency is only imposed by the PnP algorithm, which is not part of the network [22].

In this paper, we aim to address the shortcomings of current two-stage approaches. Firstly, we enforce occlusion-robust feature learning to enable models to deal with novel and severe occlusions. Secondly, a good pose representation should produce landmarks that are consistent to the object shape, rather than predicting individual landmarks independently. To this end, during model training we encourage a holistic pose representation learning in order to strengthen the connections between landmark predictions and enhance their coherence.

Our contributions We propose the Robust Object Pose Estimation (ROPE) framework which achieves excellent robustness against occlusions without the need of pose refinement. As shown in Figure 1, our model predicts landmarks and pose robustly without any post-processing.

To enforce occlusion-robust feature learning, we combine hide-and-seek [50], random erasing [65] and batch augmentation [21] and propose a occlude-and-blackout batch augmentation technique for model training. To encourage the model to learn holistic pose representations, we propose a multi-precision supervision architecture, which boosts the model's ability to extrapolate occluded object parts, leading to spatially more accurate and structurally more coherent landmark predictions. To alleviate the need for pose refinement we further utilise the multi-precision supervision architecture to filter landmark predictions with a simple verification step.

We conduct extensive experiments to verify the efficacy of the proposed techniques, and compare our method to SOTA object pose estimators. In terms of the ADD(-S) metirc, our method outperforms all contestants on LINEMOD [19] and all non-refinement methods on YCB-Video [62]. Without any refinement, it is also competitive to SOTA methods that includes a refinement step. Compared to methods that relies on large amount of synthetic training images, we show that ROPE is highly data-efficient.

2. Related works

Traditional object pose estimation methods [15, 18, 24, 19, 31, 37] rely on hand-crafted features or template matching techniques, which are susceptible to occlusions or other appearance change. Recent advancements of deep learning has nurtured a lot of learning-based methods. We briefly survey a few prominent works from one-stage, two-stage and other methods.

PoseNet [28] was a pioneer work on using a deep model to directly regress the 6DOF from an image. Although it was proposed for camera localisation rather than object pose estimation, its principle applies to both tasks. SSD-6D [27] combines an SSD detector [36] and a pose regressor in a single network. RenderForCNN [53] uses an image renderer to synthesize training images as well as groundtruth pose for training a pose regressor.

Compared to one-stage approaches, two-stage methods typically predicts intermediate features in the first stage, and then solve for the pose in the second stage. This mechanism receives more attention because its intermediate feature learning facilitates more potential improvements. For example, Tekin *et al.* [57] apply the YOLO object detector [47] in the first stage to predict object landmarks. Hu *et al.* [23] predict landmark locations for each small patch of the input image. They then aggregate all patch predictions to establish 2D-3D correspondences for solving the pose. Oberweger *et al.* [41] on the other hand, only use patches of images to train the landmark predictor. The idea is that at least some patches are not corrupted by the occluder and they could produce accurate landmark heatmaps. The ensemble of heatmaps predicted from many patches are combined to obtain final landmarks. PVNet [44] predicts the object mask and, for each pixel within the mask, unit vectors that points to the landmarks. It then utilises a generalised Hough voting scheme [2] to determine the distribution of the landmarks.

There are other notable works tackling object pose estimation from different perspectives. Sundermeyer *et al.* [56, 55] use autoencoders to learn implicit pose representations by reconstructing the input objects. Cai and Reid [5] propose a 3D model-free pose estimator via 2D-3D mapping. To make two-stage methods into a single stage pipeline, Hu *et al.* [22] and Wang *et al.* [59] propose deep architectures to replace the PnP algorithm in the second stage, while Chen *et al.* [8] propose a differentiable PnP method to achieve end-to-end learning.

3. The ROPE framework

We focus on the problem of 6DOF object pose estimation from a single RGB image. Given an image I and a known 3D point cloud $\{z_i\}_{i=1}^n$ of the target object, we first predict a set of 2D landmarks $\{x_i\}_{i=1}^n$ in I that correspond
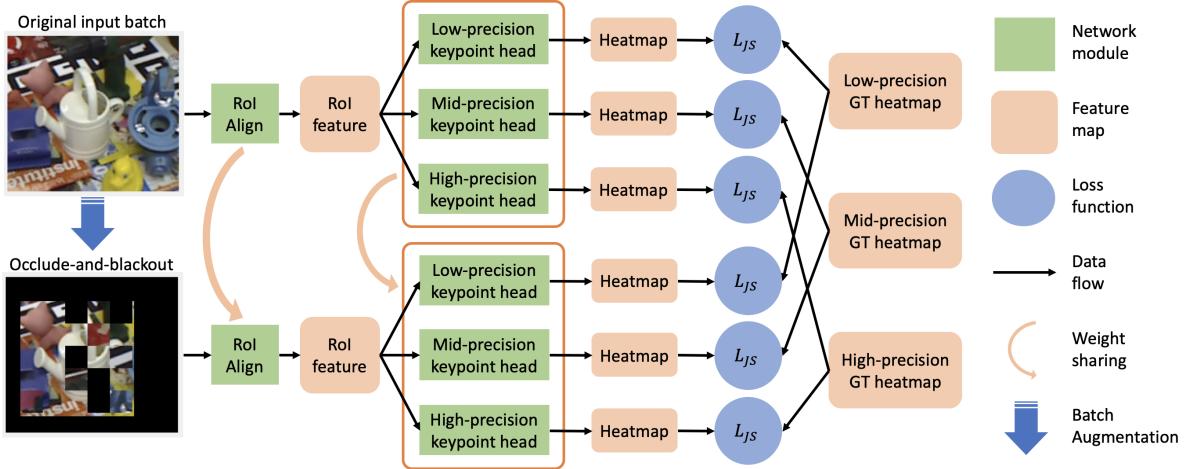
Figure 2. Illustration of an occlude-and-blackout augmented example and the architecture of our heatmap prediction network. For clarity, the backbone and the RPN are represented in the RoI Align module, other modules in the Mask R-CNN framework such as the box head, as well as relevant losses, are not shown. Our model replaces the original mask head with three keypoint heads.

to the point cloud, then solve the pose y via a RANSAC-based PnP solver from filtered 2D-3D correspondences.

3.1. Robust landmark prediction

Our 2D landmark prediction is based on the Mask R-CNN [16] framework. The specific architecture and training scheme are shown in Figure 2. A basic improvement is substituting the original backbone network with HRNet [54, 60] to exploit its high-resolution feature maps which preserve rich semantic information and increase spatial accuracy. Next, we describe two key innovations to boost occlusion invariance and landmark coherence.

3.1.1 Occlude-and-blackout batch augmentation

Fundamentally, pose estimation for the typical 3D object will suffer from the problem of self-occlusion. Landmarks that are at the opposite side of the object would be hard to predict since their visual features are hidden. In fact, a practical pose estimator must also contend with additional occlusions due to, *e.g*., other objects or scene elements that further conceal part of the target object from view. It is thus important that the landmark predictor infers the robust pose information from potentially different kinds occlusions imposed on the object.

Inspired by the ideas of random erasing [65], hide-and-seek [50], and batch augmentation [21] (all not originally developed for pose estimation), we develop a novel Occlude-and-blackout Batch Augmentation (OBA) to promote robust landmark prediction under occlusion. For each training batch, after performing regular data augmentations including rotation, translation, scaling and color jitter, we extend the batch by including a copy of itself with extra

augmentations, namely, occlude and blackout. Similar to hide-and-seek, we divide the image region enveloped by the object bounding box into a grid of patches and replace each patch, under certain probability, with either noise or a random patch elsewhere from the same image. We then blackout everything outside of the object bounding box. An example is shown in Figure 2.

With random occlusions the network is forced to infer the pose information from a partial view of the object. Erasing the background helps reducing overfitting and enhance generalisability. Moreover, the OBA augmented images are fed to the network with the original ones in the same batch, and supervised by the same groundtruth labels. This encourages the network to learn occlusion-robust and background-invariant representations.

If the potential occluders are known beforehand, injecting occluder specific information in the training phase can significantly improve performance [41]. However this knowledge is often not available in practice. Compared to methods that augment training images with known objects [25, 32, 1], our method is occluder-agnostic yet it generalises well in the testing sets.

3.1.2 Multi-precision supervision

Current heatmap-based landmark prediction networks use a single groundtruth Gaussian heatmap per landmark for training. The variance of these heatmaps is a hyper parameter which requires careful tuning: a smaller variance may increase prediction accuracy for each individual landmark however risk structural inconsistency in the case of occlusion, due to the lack of holistic understanding of the object pose. To address this issue we propose a Multi-
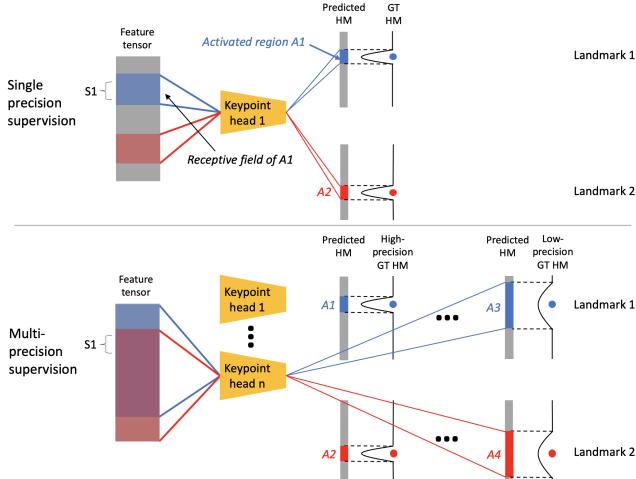
Figure 3. Conceptual illustration of holistic representation learning via MPS. Note the difference on the information learned by the feature section $S1$.

Precision Supervision (MPS) architecture: using three keypoint heads to predict groundtruth Gaussian heatmaps with different variance.

In Mask R-CNN, the output feature map of the backbone is aligned with RoI proposals and the RoI features are then passed to the mask head. We replace the mask head with three keypoint heads to regress the landmark heatmaps, as shown in Figure 2. Each keypoint head consists of 8 convolutional layers and 2 upsampling layers.

In the training phase, the groundtruth heatmaps Φ^* are constructed as 2D Gaussian feature maps centred on groundtruth 2D landmarks x^* and spreading with variance σ^2. We use σ equal to 8, 3 and 1.5 pixels respectively for the three keypoint heads, thus creating low, medium and high precision target heatmaps Φ^*. The loss function is

$$L_{JS} = \text{JSD}(\phi(\Phi), \Phi^*), \qquad (1)$$

where $\text{JSD}(\cdot)$ is the Jensen–Shannon divergence [14] and $\phi(\cdot)$ is the channel-wise softmax function, *i.e.*, each channel is normalised to be a probability distribution over the pixels.

In the testing phase, we only use the predicted heatmaps Φ from the high-precision keypoint head to obtain the landmark coordinates x. Instead of simply taking the "argmax" of Φ as x, we treat the normalised heatmaps $\phi(\Phi)$ as probability maps and take their spatial expectations as x. This has two advantages over the "argmax" approach: it has higher accuracy because it is continuous rather than discrete; it is more robust to outlying pixel values.

Although only the high-precision heatmaps are used to compute the landmark coordinates, the medium and low-precision keypoint heads play an important role in the pipeline. Firstly, having target heatmaps with different variances σ^2 helps the model adapt to objects of different sizes.

This also relieves the need for tuning σ as a hyper parameter for each object. Secondly, heatmaps from the medium-precision keypoint head are used as an auxiliary for filtering predicted landmarks, as will be explained in the next subsection. Lastly and most importantly, MPS boosts holistic representation learning in the feature maps and increases landmark coherence. An conceptual illustration is shown in Figure 3.

In Figure 3, we take one section of the feature tensor $S1$ for examination. With single precision supervision, $S1$ is only responsible for activating the region $A1$ in the predicted heatmap of Landmark 1. It does not learn useful information about Landmark 2. In the MPS scenario, besides learning about Landmark 1 via $A1$ and $A3$, $S1$ is also exposed to the receptive field of $A4$ from Landmark 2. This enforces $S1$ to incorporate relevant information and become more "aware" of the location of Landmark 2. The overall effect is that, each part of the feature tensor not only learns the necessary information to predict a local landmark, but also integrates knowledge of other landmarks to understand a wider context, thus learns a more holistic representation of the target object pose.

A holistic representation enables heatmap predictions to be more robust against occlusions. As shown in Figure 4, when trained without MPS, novel occlusions result in confused heatmap activations. On the other hand, a holistic representation learned via MPS is able to produces stable heatmaps for the occluded landmarks. This also boosts the structural consistency of landmark predictions as shown in Figure 5 and 7, which is further discussed in Section 4.3.

3.2. Landmark filtering

Many pose estimation pipelines include a refinement stage which is either optimisation-based [27, 7, 51] or learning-based [46, 34, 30, 64]. While such post-processing is effective in boosting prediction accuracy, it adds additional computation burdens which is a disadvantage especially for real-time applications. In order to boost prediction accuracy while at the same time avoiding heavy post-processing computation, we make use of the multi-heads design of MPS for selecting high-quality landmark predictions before passing them to the PnP solver, thus alleviating the need for significant pose refinement.

Specifically, for an image I, let $\{x_i\}$ denote the set of predicted landmark coordinates from the high-precision keypoint head, and $\{x_i^m\}$ denote the set of landmark coordinates predicted from the medium-precision keypoint head. We then select a subset

$$\{x_i \mid \|x_i - x_i^m\|_2 \le \epsilon\} \qquad (2)$$

for the PnP solver to compute the pose. In other words, a landmark prediction from the high-precision head will only

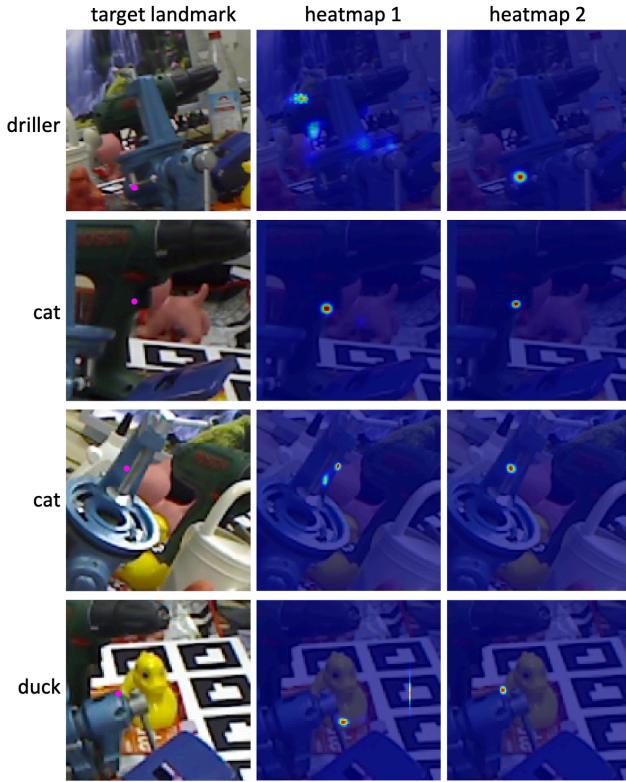| target landmark | heatmap 1 | heatmap 2 |

driller

cat

cat

duck

Figure 4. The effect of holistic representation learning in heatmap prediction. Predictions of heatmap 1 are from a model (MV1) trained without MPS while those of heatmap 2 are from the full model (original) with MPS. Details of models (MV1 and original) are provided in Section 4.3.1.

be selected for the pose solver if it is verified by the corresponding medium-precision prediction, where ϵ is the verification threshold. In the case that the selected subset has fewer than 4 points, which is the minimum number required by a PnP solver, we then use the 4 points with the smallest $\|\boldsymbol{x}_i - \boldsymbol{x}_i^m\|_2$ values as the subset.

While in this work we focus on the base pose estimator and report its performances without any refinement, our pipeline can be easily extended to stack one or multiple refiners such as [38, 33, 55].

4. Experiments

In this section we conduct experiments to validate the effectiveness of ROPE as well as to compare it to SOTA methods of RGB image-based pose estimation.

4.1. Datasets and metrics

We choose the widely used LINEMOD [19], its extension Occluded-LINEMOD [3] and the YCB-Video [62] datasets for our experiments.

For LINEMOD, we follow the convention of previous

works [46, 57, 44, 64] by using 15% of the images of each object as training set and the remaining 85% as testing set. The training images are selected in such a way that the relative rotation between them are larger than a threshold. For each object, we additionally use 1312 rendered images of the isolated object for training, which are obtained from [20]. For Occluded-LINEMOD the whole dataset is used for testing while images of the corresponding objects in LINEMOD, as well as the rendered images, are used for training. We also follow the protocol of [62, 41] for the YCB-Video dataset: we use 80 out of the 92 video sequences as well as the 80000 synthetic images for training, and test on 2949 key frames from the reserved 12 sequences.

We report the ADD(-S) metric which combines the ADD metric [19] for asymmetric objects and the ADD-S metric [62] for symmetric ones. The ADD metric computes the percentage of correctly estimated poses. A pose is considered correct if the object model points, when transformed by the predicted and groundtruth poses respectively, have an average distance of less 10% of the model diameter. For ADD-S, this distance is instead computed based on the closest point distance. The ADD(-S) metric is preferred over the 2D projection metric [4] because it directly measures the alignment discrepancy in 3D.

For the YCB-Video dataset we also report the AUC metric proposed in [62] and adopted in [41, 44]. The AUC metric is the area under the ADD(-S) curve when varying the distance threshold for a pose to be deemed correct. We vary this threshold from 0 to 10 cm, in accordance with [62].

4.2. Implementation details

For each object model we apply the farthest point sampling (FPS) algorithm [44] on the 3D point cloud and select 11 landmarks. The groundtruth 2D landmarks are then obtained by projecting the 3D landmarks with groundtruth camera pose and intrinsics. We use ImgAug [26] for regular data augmentations including rotation, translation, scaling and color jitter before the OBA. We use the Adam optimizer [29] and train the model for 250 epochs on LINEMOD and 200 epochs on Occluded-LINEMOD and YCB-Video. We set the landmark verification threshold ϵ to 1 pixel for all datasets.

4.3. Ablation studies

We conduct various ablation tests to investigate the effect of the proposed OBA and MPS.

4.3.1 Model variations

To verify the efficacy of OBA and MPS, we create two Model Variants (MV) of ROPE:

Figure 5. Comparing performances of model variants on the Occluded-LINEMOD dataset with qualitative examples.
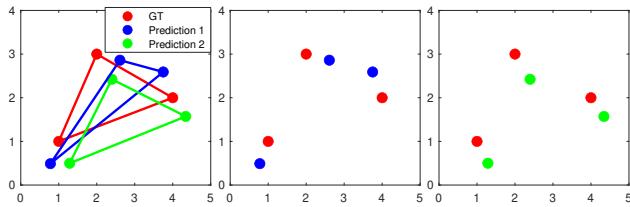

Figure 6. A toy example for the intuition of incoherence measure c_i. The mean residual r_i for prediction 1 (blue) and prediction 2 (green) are both 0.608. However, their mean incoherence measure c_i are 0.604 and 0.074, respectively. Although both predictions are identical in terms of accuracy, prediction 2 has much better coherence as the green triangle is much more similar in shape to the groundtruth than the blue one.

1. (MV1: w/ OBA, w/o MPS) While keeping everything else of the original ROPE unchanged, we remove the low and medium-precision keypoint heads, and train the one-head-model with high-precision groundtruth heatmaps.

2. (MV2: w/o OBA, w/o MPS) On top MV1, we further remove OBA in training. Note that *common data augmentations including rotation, translation, scaling and color jitter, are still kept*.

Figure 5 shows the overall ADD(-S) on the Occluded-LINEMOD dataset, as well as qualitative results of all model variants. Without both OBA and MPS, object detection can easily fail and landmark prediction is precarious. We can clearly see that occlusion-robust feature learning enforced by OBA significantly increases the reliability of object detection and landmark prediction. In addition, by comparing MV1 and the original model, it is obvious that MPS boosts the structural consistency of the predicted landmarks, especially in occluded regions. This shows that a holistic representation induced by MPS enhances landmark

Figure 7. Comparing the results of training with and without MPS on LINEMOD, while keeping all else equal. The vertical location of each bubble represents the mean prediction residual r_i of all landmarks in the testing sets. The size of each bubble indicates the mean incoherence c_i.

coherence, strengthening the model's ability to extrapolate to the occluded part of the object.

4.3.2 Accuracy and coherence of landmarks

To formally analyse the effect of holistic representation learning, we quantify accuracy and structural consistency of landmark predictions and compare them when trained with and without MPS. For accuracy, we define

$$r_i = \|x_i - x_i^*\|_2 \tag{3}$$

as the prediction residual of a 2D landmark x_i. We also define a measure of incoherence

$$c_i = \|(x_i - x_i^*) - m\|_2 \tag{4}$$

ADD(-S)	Without refinement						With refinement			
	PVNet [44]	Pix2Pose [42]	DPOD [64]	CDPN [35]	GDR [59]	Ours	SSD-6D [27]	DPOD+ [64]	HybridPose [51]	DeepIM [33]
ape	43.62	58.10	53.28	64.38	-	**81.52**	65.00	**87.70**	63.10	77.00
benchevise	99.90	91.00	95.34	97.77	-	**100.00**	80.00	98.50	**99.90**	97.50
cam	86.86	60.90	90.36	91.67	-	**96.86**	78.00	**96.10**	90.40	93.50
can	95.47	84.40	94.10	95.87	-	**98.72**	86.00	**99.70**	98.50	96.50
cat	79.34	65.00	60.38	83.83	-	**94.71**	70.00	**94.70**	89.40	82.10
driller	96.43	76.30	97.72	96.23	-	**99.01**	73.00	**98.80**	98.50	95.00
duck	52.58	43.80	66.01	66.76	-	**85.35**	66.00	**86.30**	65.00	77.70
eggbox*	99.15	96.80	99.72	99.72	-	**100.00**	100.00	99.90	**100.00**	97.10
glue*	95.66	79.40	93.83	**99.61**	-	99.42	100.00	96.80	98.80	99.40
holepuncher	81.92	74.80	65.83	85.82	-	**90.39**	49.00	86.90	**89.70**	52.80
iron	98.88	83.40	99.80	97.85	-	**100.00**	78.00	**100.00**	100.00	98.30
lamp	99.33	82.00	88.11	97.89	-	**99.42**	73.00	96.80	**99.50**	97.50
phone	92.41	45.00	74.24	90.75	-	**97.64**	79.00	94.70	**94.90**	87.70
average	86.27	72.38	82.98	89.86	93.70	**95.61**	76.69	**95.15**	91.36	88.60

Table 1. Test accuracy on the LINEMOD dataset in terms of the ADD(-S) metric. Objects with a "*" sign are considered as symmetric objects and the ADD-S metric is used. The result of SSD-6D is obtained from [57]. The result of HybridPose is from its fourth version update in [52].

ADD(-S)	Without refinement								With refinement	
	HM [41]	PVNet [44]	Hu [23]	Pix2Pose [42]	DPOD [64]	Hu2 [22]	GDR [59]	Ours	DPOD+ [64]	HybridPose [51]
ape	15.30	15.81	12.10	22.00	-	19.20	**39.30**	28.03	-	**20.90**
can	44.70	63.30	39.90	44.70	-	65.10	**79.20**	75.06	-	**75.30**
cat	9.33	16.68	8.20	22.70	-	18.90	23.50	**25.53**	-	**24.90**
driller	55.40	65.65	45.20	44.70	-	69.00	**71.30**	61.86	-	**70.20**
duck	19.60	25.24	17.20	15.00	-	25.30	**44.40**	19.07	-	**27.90**
eggbox*	23.00	50.17	22.10	25.20	-	52.00	**58.20**	45.62	-	**52.40**
glue*	41.40	49.62	35.80	32.40	-	51.40	49.30	**56.92**	-	**53.80**
holepuncher	20.40	39.67	36.00	49.50	-	45.60	**58.70**	55.54	-	**54.20**
average	28.64	40.77	27.06	32.03	32.80	43.30	**53.00**	45.95	47.30	**47.45**

Table 2. Test accuracy on the Occluded-LINEMOD dataset in terms of the ADD(-S) metric. Objects with a "*" sign are considered as symmetric objects and the ADD-S metric is used. The result of HybridPose is from its fourth version update in [52].

for a landmark prediction x_i where $m = \frac{1}{n}\sum_{i=1}^{n}(x_i - x_i^*)$ is the mean error vector for an image. The smaller c_i is, the more coherent a prediction x_i is, resulting a more consistent structure of prediction to the groundtruth. An intuitive example is shown in Figure 6.

As shown in Figure 7, training with MPS effectively lowers the mean residuals. Furthermore, the mean incoherence are also smaller for all objects. This confirms that a more holistic understanding of the object pose can produce more accurate and structurally consistent landmark predictions.

4.4. Comparing to SOTA methods

We report results on the LINEMOD dataset in Table 1. We group methods into two types depending on whether they include a separate refinement step or not. Our method achieves the best average ADD(-S), as well as the best ADD(-S) on most individual objects. Moreover, our method even outperforms all SOTA methods with refinement, further attesting the power of ROPE. The results on the Occluded-LINEMOD dataset are summarised in Table 2. In the non-refinement group, our method ranked second amongst current SOTA methods overall and best on two individual objects. A sample of qualitative results are provided in Figures 1 and 5. The results on the YCB-Video dataset are reported in Table 3. Without refinement, ROPE has the best performance when evaluated with ADD(-S).

4.5. Data efficiency

The LINEMOD dataset has about 1200 images for each object, which results in approximately 180 images (15%)

	ADD(-S)					AUC of ADD(-S)					
	Without refinement					Without refinement				With refinement	
	HM [41]	Hu [23]	Hu2 [22]	GDR [59]	Ours	HM [41]	PVNet [44]	GDR [59]	Ours	DeepIM [34]	CosyPose [30]
master chef can	31.20	33.00	-	-	**46.52**	69.00	-	-	**71.17**	**71.20**	-
cracker box	75.00	44.60	-	-	**92.63**	80.20	-	-	**89.86**	83.60	-
sugar box	47.20	75.60	-	-	**99.15**	76.20	-	-	**93.21**	94.10	-
tomato soup can	30.20	40.80	-	-	**60.90**	70.00	-	-	**82.53**	86.10	-
mustard bottle	72.50	70.60	-	-	**100.00**	84.80	-	-	**95.34**	91.50	-
tuna fish can	4.31	18.10	-	-	**52.96**	49.40	-	-	**88.01**	87.70	-
pudding box	48.30	12.20	-	-	**79.91**	82.20	-	-	**90.5**	82.70	-
gelatin box	37.20	**59.40**	-	-	58.88	81.80	-	-	**89.36**	91.90	-
potted meat can	40.30	33.30	-	-	**58.62**	66.20	-	-	**74.54**	76.20	-
banana	6.20	16.60	-	-	**36.94**	52.90	-	-	**58.77**	81.20	-
pitcher base	53.80	90.00	-	-	**99.65**	69.90	-	-	**92.86**	90.10	-
bleach cleanser	57.20	70.90	-	-	**75.22**	73.30	-	-	**77.35**	81.20	-
bowl*	**49.50**	30.50	-	-	45.07	**80.30**	-	-	70.81	81.40	-
mug	10.50	40.70	-	-	**66.04**	50.50	-	-	**89.1**	81.40	-
power drill	63.00	63.50	-	-	**94.99**	78.30	-	-	**89.4**	85.50	-
wood block*	48.20	27.70	-	-	**55.37**	65.20	-	-	**70.62**	81.90	-
scissors	0.55	17.10	-	-	**71.27**	28.20	-	-	**84.82**	60.90	-
large marker	11.70	4.80	-	-	**11.73**	48.20	-	-	**53.25**	75.60	-
large clamp*	12.20	25.60	-	-	**68.12**	47.20	-	-	**77.1**	74.30	-
extra large clamp*	17.30	8.80	-	-	**56.16**	47.50	-	-	**55.19**	73.30	-
foam brick*	63.80	34.70	-	-	**68.40**	85.60	-	-	83.78	81.90	-
average	37.15	38.98	53.90	60.10	**66.59**	66.04	73.40	**84.40**	79.88	81.90	**84.50**

Table 3. Test accuracy on the YCB-Video dataset. Objects with a "*" sign are considered as symmetric objects.

for the training set. To supplement such a small training set many methods generate a large amount of synthetic images. For example, PVNet [44] renders 20000 images for each object and the same strategy is adopted in [52]. Although we only use a moderate amount of 1312 synthetic images on top of the 180 in training, we test our model's performance in a extremely data-efficient case: only using the ~180 images for training.

As shown in Table 4, despite having slightly lower ADD(-S) then the baseline, our model achieves an overall accuracy of 93.22% which is close to the current SOTA method GDR [59]. This is accomplished with as few as around 180 training images, demonstrating superior data efficiency for our method.

5. Conclusion

We propose ROPE, a framework for robust object pose estimation against occlusions. We show that enforcing occlusion-robust feature learning and encouraging holistic representation learning are the key to achieve occlusion-robustness. Evaluations on three popularly used benchmark datasets, LINEMOD, Occluded-LINEMOD and YCB-Video, show that ROPE either outperforms or is com-

ADD(-S)	Training images	
	~180	~1500
ape	78.57	**81.52**
benchvise	98.93	**100.00**
cam	90.88	**96.86**
can	98.03	**98.72**
cat	92.22	**94.71**
driller	98.02	**99.01**
duck	79.06	**85.35**
eggbox*	99.72	**100.00**
glue*	97.68	**99.42**
holepuncher	88.30	**90.39**
iron	96.83	**100.00**
lamp	98.85	**99.42**
phone	94.81	**97.64**
average	93.22	**95.61**

Table 4. Comparing performances of ROPE in the extremely data-efficient setting (~180) and in the original setting (~1500) on the LINEMOD dataset. Both models are without refinement.

petitive to SOTA methods, without the need of refinement. Our method is also highly data-efficient.

References

[1] Hassan Abu Alhaija, Siva Karthik Mustikovela, Lars Mescheder, Andreas Geiger, and Carsten Rother. Augmented reality meets deep learning for car instance segmentation in urban scenes. In *BMVC*, 2017. 3

[2] Dana H Ballard. Generalizing the hough transform to detect arbitrary shapes. *Pattern recognition*, 13(2):111–122, 1981. 2

[3] Eric Brachmann, Alexander Krull, Frank Michel, Stefan Gumhold, Jamie Shotton, and Carsten Rother. Learning 6d object pose estimation using 3d object coordinates. In *ECCV*, 2014. 5

[4] Eric Brachmann, Frank Michel, Alexander Krull, Michael Ying Yang, Stefan Gumhold, and Carsten Rother. Uncertainty-driven 6d pose estimation of objects and scenes from a single rgb image. In *CVPR*, 2016. 5

[5] Ming Cai and Ian Reid. Reconstruct locally, localize globally: A model free method for object pose estimation. In *CVPR*, 2020. 2

[6] Lorenzo Pasqualetto Cassinis, Robert Fonod, and Eberhard Gill. Review of the robustness and applicability of monocular pose estimation systems for relative navigation with an uncooperative spacecraft. *Progress in Aerospace Sciences*, 2019. 1

[7] Bo Chen, Jiewei Cao, Alvaro Parra, and Tat-Jun Chin. Satellite pose estimation with deep landmark regression and nonlinear pose refinement. In *ICCVW*, 2019. 4

[8] Bo Chen, Alvaro Parra, Jiewei Cao, Nan Li, and Tat-Jun Chin. End-to-end learnable geometric vision by backpropagating pnp optimization. In *CVPR*, 2020. 2

[9] Wei Chen, Xi Jia, Hyung Jin Chang, Jinming Duan, Linlin Shen, and Ales Leonardis. Fs-net: Fast shape-based network for category-level 6d object pose estimation with decoupled rotation mechanism. In *CVPR*, pages 1581–1590, 2021. 1

[10] Xiaozhi Chen, Huimin Ma, Ji Wan, Bo Li, and Tian Xia. Multi-view 3d object detection network for autonomous driving. In *CVPR*, 2017. 1

[11] Alvaro Collet, Manuel Martinez, and Siddhartha S Srinivasa. The moped framework: Object recognition and pose estimation for manipulation. *The international journal of robotics research*, 30(10):1284–1306, 2011. 1

[12] A. Crivellaro, M. Rad, Y. Verdie, K. M. Yi, P. Fua, and V. Lepetit. Robust 3d object tracking from monocular images using stable parts. *TPAMI*, 40(6):1465–1479, 2018. 1

[13] Thanh-Toan Do, Ming Cai, Trung Pham, and Ian Reid. Deep-6dpose: Recovering 6d object pose from a single rgb image. *arXiv preprint arXiv:1802.10367*, 2018. 1

[14] Bent Fuglede and Flemming Topsoe. Jensen-shannon divergence and hilbert space embedding. In *Proceedings of The International Symposium on Information Theory*, 2004. 4

[15] Chunhui Gu and Xiaofeng Ren. Discriminative mixture-of-templates for viewpoint classification. In *ECCV*, 2010. 2

[16] Kaiming He, Georgia Gkioxari, Piotr Dollar, and Ross Girshick. Mask r-cnn. *TPAMI*, 2018. 3

[17] Yisheng He, Wei Sun, Haibin Huang, Jianran Liu, Haoqiang Fan, and Jian Sun. Pvn3d: A deep point-wise 3d keypoints voting network for 6dof pose estimation. In *CVPR*, 2020. 1

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

[19] Stefan Hinterstoisser, Vincent Lepetit, Slobodan Ilic, Stefan Holzer, Gary Bradski, Kurt Konolige, and Nassir Navab. Model based training, detection and pose estimation of texture-less 3d objects in heavily cluttered scenes. In *ACCV*, 2012. 2, 5

[20] Tomas Hodan, Frank Michel, Eric Brachmann, Wadim Kehl, Anders GlentBuch, Dirk Kraft, Bertram Drost, Joel Vidal, Stephan Ihrke, Xenophon Zabulis, et al. Bop: Benchmark for 6d object pose estimation. In *ECCV*, 2018. 5

[21] Elad Hoffer, Tal Ben-Nun, Itay Hubara, Niv Giladi, Torsten Hoefler, and Daniel Soudry. Augment your batch: Improving generalization through instance repetition. In *CVPR*, 2020. 2, 3

[22] Yinlin Hu, Pascal Fua, Wei Wang, and Mathieu Salzmann. Single-stage 6d object pose estimation. In *CVPR*, 2020. 1, 2, 7, 8

[23] Yinlin Hu, Joachim Hugonot, Pascal Fua, and Mathieu Salzmann. Segmentation-driven 6d object pose estimation. In *CVPR*, 2019. 1, 2, 7, 8

[24] Daniel P Huttenlocher, Gregory A. Klanderman, and William J Rucklidge. Comparing images using the hausdorff distance. *TPAMI*, 15(9):850–863, 1993. 2

[25] Omid Hosseini Jafari, Siva Karthik Mustikovela, Karl Pertsch, Eric Brachmann, and Carsten Rother. ipose: instance-aware 6d pose estimation of partly occluded objects. In *ACCV*, 2018. 1, 2, 3

[26] Alexander B. Jung, Kentaro Wada, Jon Crall, Satoshi Tanaka, Jake Graving, Christoph Reinders, Sarthak Yadav, Joy Banerjee, Gábor Vecsei, Adam Kraft, Zheng Rui, Jirka Borovec, Christian Vallentin, Semen Zhydenko, Kilian Pfeiffer, Ben Cook, Ismael Fernández, François-Michel De Rainville, Chi-Hung Weng, Abner Ayala-Acevedo, Raphael Meudec, Matias Laporte, et al. imgaug. https://github.com/aleju/imgaug, 2020. Online; accessed 01-Feb-2020. 5

[27] Wadim Kehl, Fabian Manhardt, Federico Tombari, Slobodan Ilic, and Nassir Navab. Ssd-6d: Making rgb-based 3d detection and 6d pose estimation great again. In *ICCV*, 2017. 1, 2, 4, 7

[28] Alex Kendall, Matthew Grimes, and Roberto Cipolla. Posenet: A convolutional network for real-time 6-dof camera relocalization. In *ICCV*, 2015. 2

[29] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 5

[30] Yann Labbé, Justin Carpentier, Mathieu Aubry, and Josef Sivic. Cosypose: Consistent multi-view multi-object 6d pose estimation. In *ECCV*, 2020. 4, 8

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

[32] Chi Li, M Zeeshan Zia, Quoc-Huy Tran, Xiang Yu, Gregory D Hager, and Manmohan Chandraker. Deep supervi-

sion with shape concepts for occlusion-aware 3d object parsing. In *CVPR*, 2017. 3

[33] Yi Li, Gu Wang, Xiangyang Ji, Yu Xiang, and Dieter Fox. Deepim: Deep iterative matching for 6d pose estimation. In *ECCV*, 2018. 5, 7

[34] Yi Li, Gu Wang, Xiangyang Ji, Yu Xiang, and Dieter Fox. Deepim: Deep iterative matching for 6d pose estimation. *International Journal of Computer Vision*, 128(3):657–678, 2020. 4, 8

[35] Zhigang Li, Gu Wang, and Xiangyang Ji. Cdpn: Coordinates-based disentangled pose network for real-time rgb-based 6-dof object pose estimation. In *ICCV*, 2019. 1, 2, 7

[36] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *ECCV*, 2016. 2

[37] David G Lowe. Object recognition from local scale-invariant features. In *ICCV*, 1999. 2

[38] Fabian Manhardt, Wadim Kehl, Nassir Navab, and Federico Tombari. Deep model-based 6d pose refinement in rgb. In *ECCV*, 2018. 5

[39] Eric Marchand, Hideaki Uchiyama, and Fabien Spindler. Pose estimation for augmented reality: a hands-on survey. *IEEE transactions on visualization and computer graphics*, 22(12):2633–2651, 2015. 1

[40] Frank Michel, Alexander Kirillov, Eric Brachmann, Alexander Krull, Stefan Gumhold, Bogdan Savchynskyy, and Carsten Rother. Global hypothesis generation for 6d object pose estimation. In *CVPR*, 2017. 1

[41] Markus Oberweger, Mahdi Rad, and Vincent Lepetit. Making deep heatmaps robust to partial occlusions for 3d object pose estimation. In *ECCV*, 2018. 1, 2, 3, 5, 7, 8

[42] Kiru Park, Timothy Patten, and Markus Vincze. Pix2pose: Pixel-wise coordinate regression of objects for 6d pose estimation. In *ICCV*, 2019. 1, 2, 7

[43] Georgios Pavlakos, Xiaowei Zhou, Aaron Chan, Konstantinos G Derpanis, and Kostas Daniilidis. 6-dof object pose from semantic keypoints. In *ICRA*, 2017. 1

[44] Sida Peng, Yuan Liu, Qixing Huang, Xiaowei Zhou, and Hujun Bao. Pvnet: Pixel-wise voting network for 6dof pose estimation. In *CVPR*, 2019. 1, 2, 5, 7, 8

[45] Patrick Poirson, Phil Ammirato, Cheng-Yang Fu, Wei Liu, Jana Kosecka, and Alexander C Berg. Fast single shot detection and pose estimation. In *The Fourth International Conference on 3D Vision*. IEEE, 2016. 1

[46] Mahdi Rad and Vincent Lepetit. Bb8: A scalable, accurate, robust to partial occlusion method for predicting the 3d poses of challenging objects without using depth. In *ICCV*, 2017. 1, 4, 5

[47] Joseph Redmon and Ali Farhadi. Yolo9000: better, faster, stronger. In *CVPR*, 2017. 2

[48] Torsten Sattler, Qunjie Zhou, Marc Pollefeys, and Laura Leal-Taixe. Understanding the limitations of cnn-based absolute camera pose regression. In *CVPR*, 2019. 1

[49] S. Sharma, C. Beierle, and S. D'Amico. Pose estimation for non-cooperative spacecraft rendezvous using convolutional neural networks. In *IEEE Aerospace Conference*, 2018. 1

[50] Krishna Kumar Singh and Yong Jae Lee. Hide-and-seek: Forcing a network to be meticulous for weakly-supervised object and action localization. In *ICCV*, 2017. 2, 3

[51] Chen Song, Jiaru Song, and Qixing Huang. Hybridpose: 6d object pose estimation under hybrid representations. In *CVPR*, 2020. 1, 4, 7

[52] Chen Song, Jiaru Song, and Qixing Huang. Hybridpose: 6d object pose estimation under hybrid representations. *arXiv preprint arXiv:2001.01869*, 2020. 7, 8

[53] Hao Su, Charles R Qi, Yangyan Li, and Leonidas J Guibas. Render for cnn: Viewpoint estimation in images using cnns trained with rendered 3d model views. In *ICCV*, 2015. 2

[54] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *CVPR*, 2019. 3

[55] Martin Sundermeyer, Maximilian Durner, En Yen Puang, Zoltan-Csaba Marton, Narunas Vaskevicius, Kai O Arras, and Rudolph Triebel. Multi-path learning for object pose estimation across domains. In *CVPR*, 2020. 2, 5

[56] Martin Sundermeyer, Zoltan-Csaba Marton, Maximilian Durner, and Rudolph Triebel. Augmented autoencoders: Implicit 3d orientation learning for 6d object detection. *IJCV*, 128(3):714–729, 2020. 2

[57] Bugra Tekin, Sudipta N Sinha, and Pascal Fua. Real-time seamless single shot 6d object pose prediction. In *CVPR*, 2018. 1, 2, 5, 7

[58] Chen Wang, Danfei Xu, Yuke Zhu, Roberto Martín-Martín, Cewu Lu, Li Fei-Fei, and Silvio Savarese. Densefusion: 6d object pose estimation by iterative dense fusion. In *CVPR*, 2019. 1

[59] Gu Wang, Fabian Manhardt, Federico Tombari, and Xiangyang Ji. Gdr-net: Geometry-guided direct regression network for monocular 6d object pose estimation. In *CVPR*, 2021. 2, 7, 8

[60] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang, W. Liu, and B. Xiao. Deep high-resolution representation learning for visual recognition. *TPAMI*, 2020. 3

[61] Di Wu, Zhaoyong Zhuang, Canqun Xiang, Wenbin Zou, and Xia Li. 6d-vnet: End-to-end 6-dof vehicle pose estimation from monocular rgb images. In *CVPRW*, 2019. 1

[62] Yu Xiang, Tanner Schmidt, Venkatraman Narayanan, and Dieter Fox. Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes. In *Robotics: Science and Systems (RSS)*, 2018. 1, 2, 5

[63] Danfei Xu, Dragomir Anguelov, and Ashesh Jain. Pointfusion: Deep sensor fusion for 3d bounding box estimation. In *CVPR*, 2018. 1

[64] Sergey Zakharov, Ivan Shugurov, and Slobodan Ilic. Dpod: 6d pose object detector and refiner. In *ICCV*, 2019. 1, 4, 5, 7

[65] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. In *AAAI*, 2020. 2, 3

[66] Menglong Zhu, Konstantinos G Derpanis, Yinfei Yang, Samarth Brahmbhatt, Mabel Zhang, Cody Phillips, Matthieu Lecce, and Kostas Daniilidis. Single image 3d object detection and pose estimation for grasping. In *ICRA*, 2014. 1

[67] Yiming Zuo, Weichao Qiu, Lingxi Xie, Fangwei Zhong, Yizhou Wang, and Alan L. Yuille. Craves: Controlling robotic arm with a vision-based economic system. In *CVPR*, 2019. 1