Rigidity-Aware Detection for 6D Object Pose Estimation

Yang Hai [1], Rui Song [1], Jiaojiao Li [1], Mathieu Salzmann [2,3], Yinlin Hu [4]
[1] Xidian University, [2] EPFL, [3] ClearSpace, [4] MagicLeap
yanghai1218@gmail.com, {rsong, jjli}@xidian.edu.cn,
mathieu.salzmann@epfl.ch, yhu@magicleap.com

Abstract

Most recent 6D object pose estimation methods first use object detection to obtain 2D bounding boxes before actually regressing the pose. However, the general object detection methods they use are ill-suited to handle cluttered scenes, thus producing poor initialization to the subsequent pose network. To address this, we propose a rigidity-aware detection method exploiting the fact that, in 6D pose estimation, the target objects are rigid. This lets us introduce an approach to sampling positive object regions from the entire visible object area during training, instead of naively drawing samples from the bounding box center where the object might be occluded. As such, every visible object part can contribute to the final bounding box prediction, yielding better detection robustness. Key to the success of our approach is a visibility map, which we propose to build using a minimum barrier distance between every pixel in the bounding box and the box boundary. Our results on seven challenging 6D pose estimation datasets evidence that our method outperforms general detection frameworks by a large margin. Furthermore, combined with a pose regression network, we obtain state-of-the-art pose estimation results on the challenging BOP benchmark.

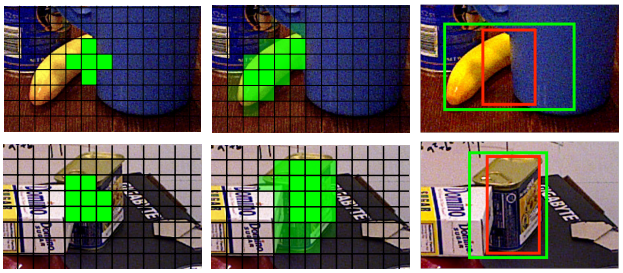
1. Introduction

Estimating the 6D pose of objects, i.e., their 3D rotation and 3D translation w.r.t. the camera, is a fundamental computer vision problem with many applications in, e.g., robotics, quality control, and augmented reality. Most recent methods [3,6,10,25,42,45] follow a two-stage pipeline: First, they detect the objects, and then estimate their 6D pose from a resized version of the resulting detected image patches. While this approach works well in simple scenarios, its performance drops significantly in the presence of cluttered scenes. In particular, and as illustrated in Fig. 1, we observed this to be mainly caused by detection failures.

Specifically, most 6D pose estimation methods rely on standard object detection methods [11, 22, 37, 43, 44, 50], which were designed to handle significantly different scenes than those observed in 6D object pose estimation benchmarks, typically with much smaller occlusions, as shown in Fig. 1(a). Because of these smaller occlusions, standard detection methods make the assumption that the regions in the center of the ground-truth bounding boxes depict the object of interest, and thus focus on learning to predict the bound-

(a) General detection (b) Detection in 6D Pose
(c) Baseline detection results (d) Our detection results
(e) Baseline pose results (f) Our pose results

Figure 1. **The challenges of detection in 6D object pose. (a)** The general detection scenario (COCO [29]) exhibits small occlusions. **(b)** The occlusion problem in 6D object pose, however, is much more severe, **(c)** making the general detection method [44] based on center-oriented sampling unreliable (glue) or fail completely (cat). **(d)** By contrast, our new detection strategy is effective in these challenging scenarios, **(e,f)** and provides significantly more robust 2D box initialization for the following 6D regression networks [15], yielding more accurate pose estimates.



(a) Baseline strategy (b) Our strategy (c) Detection results

Figure 2. **Detecting rigid objects in cluttered scenes. (a)** The standard strategy [50] chooses positive samples (green cells) around the object center, thus suffering from occlusions. **(b)** Instead, we propose to use a visibility-guided sampling strategy to discard the occluded regions and encourage the network to be supervied by all visible parts. The sampling probability is depicted by different shades of green. (**(c)** Our method (green boxes) yields more accurate detections than the standard strategy [50] (red boxes).

ing box parameters from samples drawn from these regions only. However, as shown in Fig. 2, this is ill-suited to 6D pose estimation in cluttered scenes, where the center of the objects is often occluded by other objects or scene elements.

To handle this, we propose a detection approach that leverages the property that the target objects in 6D pose estimation are rigid. For such objects, any visible parts can provide a reliable estimate of the complete bounding box. We therefore argue that, in contrast with the center-based sampling used by standard object detectors, any, and only feature vectors extracted from the visible parts should be potential candidates of positive samples during training.

In principle, modeling the visibility could be achieved by annotating segmentation masks for all objects. This process, however, is cumbersome, particularly in the presence of occlusions by scene elements, and would limit the scalability of the approach. Instead, we therefore propose to compute a visibility probability based on a minimum barrier distance between any pixel in a bounding box and the box boundary. We then use this probability to guide the sampling of candidates during training, thus discarding the occluded regions and encouraging the network to be supervised by all visible parts. Furthermore, to leverage the reliability of local predictions from most visible parts during inference, we collect all candidate local predictions above a confidence threshold, and combine them by a simple weighted average, yielding more robust detections.

We demonstrate the effectiveness of our method on seven challenging 6D object pose estimation datasets, on which we consistently and significantly outperform all detection baselines. Furthermore, combined with a 6D pose regression network, our approach yields state-of-the-art results.

2. Related Work

Object pose estimation, whose goal is to estimate the 3D rotation and 3D translation of a target object with respect to the camera, nowadays typically involves a pose regression network to establish 3D-to-2D correspondences [12, 15–17, 19, 20, 33]. These correspondences then act as input to a perspective-n-points solver (PnP) [24] to compute the final 6D object pose. The current state-of-the-art methods [3, 6, 10, 23, 26, 32, 40, 42, 46] virtually all use a 2D object detector to allow the following pose regression networks to focus on a region of interest (RoI), thus yielding more accurate poses.

While this is effective when detection is successful, the pose accuracy deteriorates significantly in case of missing or inaccurate detections. In particular, 6D pose estimation frameworks typically use standard object detectors that, as shown in Figs. 1 and 2, often fail in cluttered scenes such as those of standard 6D pose estimation benchmarks as they were not designed to handle such situations. To handle this, we propose a rigidity-aware detection method that leverages the target properties. As shown by our results, it yields significant better RoIs for 6D object pose estimation.

Object detection, whose goal is to extract accurate 2D bounding boxes for all objects in a scene, has been widely studied in 2D computer vision. Existing methods follow one of two main strategies: two-stage or one-stage detection. Two-stage detectors first employ a region proposal network [11, 37] to generate bounding box candidates, which are then processed by a classification and refinement network to remove false positives and adjust the bounding boxes position and size [4, 11, 37]. Although this strategy is accurate in general, it is costly and inefficient in practice.

One-stage detectors tackle this by replacing the region proposal network with a pre-defined set of anchors at every spatial location in the encoder's final feature map [28, 34, 43]. Unfortunately, this suffers from the presence of many negative samples among the anchors. While this can be addressed to some degree by the focal loss of [27, 28], early single-stage detectors did not reach the accuracy of two-stage ones.

This was addressed in [50] via a simple yet effective strategy to sample positive candidates in a one-stage detector. Most recent detection methods follow similar strategies [9, 22, 31, 35, 36, 44, 52], and now achieve better accuracy than two-stage methods while being more efficient.

Nevertheless, while these methods work well on standard object detection benchmarks, they suffer from the heavy occlusions present in 6D pose estimation ones. Here, we therefore propose a new strategy dedicated to detecting rigid objects, and show that it outperforms standard detectors by a large margin in our 6D pose estimation context.

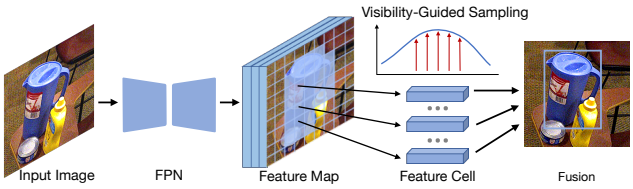
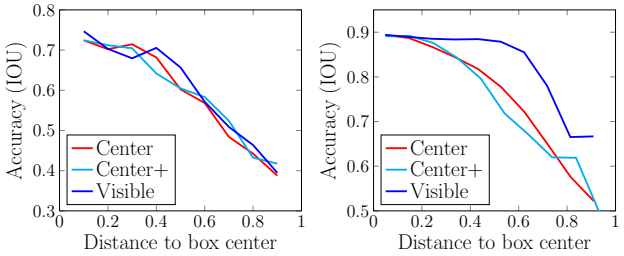


Figure 3. **Overview of our detection approach.** We use a general Feature Pyramid Network (FPN) as our backbone. We first compute a probability of visibility for every local area within the bounding box, which we use to guide the sampling of positive cells during training, without any mask annotations. Finally, during inference, we combine all the local candidate predictions to obtain a more robust final result.



(a) General (COCO) (b) Rigid (YCB-V)

Figure 4. **Analysis of rigidity in detection.** We show the testing accuracy of different sampling strategies w.r.t. different local predictions during training on a typical general object dataset (COCO [29]) and on a typical 6D object pose dataset (YCB-V [48]). We report the results of FCOSv2 [44] (Center), ATSS [50] (Center+), and a strategy exploiting all the candidates in the ground-truth mask (Visible). The horizontal axis represents the normalized distance of a local prediction to the box center. Although the accuracy of different strategies is similar on COCO, the visibility-guided sampling is much more accurate on YCB-V, even when the local predictions come from non-center areas, thanks to the rigidity of the target objects.

3. Approach

Given an RGB image depicting rigid objects, our goal is to estimate the 2D bounding box of each potential target for the subsequent pose regression network. To address this, we propose to leverage the fact that, in the context of 6D object pose estimation, we observe rigid targets. In this section, we first briefly review the problem of positive sampling in object detection and analyze the influence of the objects' rigidity in 6D object pose scenarios. We then explain how we compute object foreground probabilities without having access to ground-truth masks, and introduce a positive sampling strategy based on these probabilities. Finally, we propose a box fusion strategy to improve detection robustness. Fig. 3 provides an overview of our detection approach.

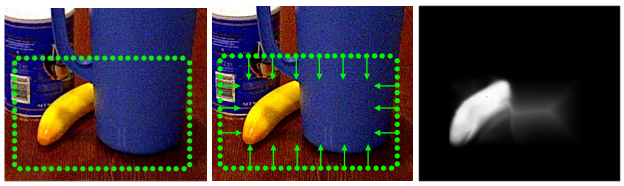
3.1. Analysis of Rigidity in Detection

Modern single-stage object detectors [22, 28, 43, 50, 52] rely on a Feature Pyramid Network (FPN) [27] that outputs scale-rich feature maps. Each feature vector is taken as a training sample and further processed by a classification branch and a regression branch. Training the detector thus first requires defining positive and negative samples for each annotated object instance. The positive samples are then encouraged to be classified as the instance's category, whereas the negative samples should be predicted as background. Furthermore, the positive samples should regress the instance's bounding box parameters. Since during training a single instance is associated with multiple positive samples, at inference multiple samples will be activated for a potential target. Most methods then use the standard Non-Maximum Suppression (NMS) as a post-processing stage to obtain the final result.

Key to the success of this general framework is the selection of positive samples during training. The standard approach to sampling positive features during training consists of assuming that the regions in the center of the ground-truth bounding boxes depict the object. However, in the context of 6D pose estimation, this center assumption is often violated because of the large occlusions that occur in cluttered scenes. More importantly, it does not account for the fact that, for rigid target objects, all visible object parts can provide a reliable prediction of the entire bounding box.

To evidence this, we train the same FPN network with different sampling strategies on the general COCO dataset [29] dataset and on the typical 6D object pose YCB-V [48] dataset, respectively. We first evaluate two baseline strategies, consisting of sampling a fixed number of positive cells from the center region in the ground-truth bounding box (FCOSv2 [44]), and of an adaptive center-based sampling strategy across all pyramid feature levels (ATSS [50]). Furthermore, we evaluate a sampling strategy that randomly chooses 10 positive cells within the ground-truth object mask.

Fig. 4 depicts the average test accuracy of different local predictions obtained with these sampling strategies as a function of the distance of the prediction to the true bounding box center. On the general COCO dataset, the accuracy deteriorates as the distance increases regardless of which sampling strategies was used during training. This comes from the diversity of the object types in COCO, which includes many non-rigid objects and a wide variety of instances with the same object type, making the object center a more reliable predictor of the bounding box. On YCB-V, the accuracy of the centered-based strategies also deteriorates quickly as the distance increases, since most non-center area were not involved during training. However, thanks to the rigidity of the YCB-V targets, the visibility-guided strategy yields more accurate local predictions, even for those that are farther away from the center area.



(a) Initial seeds (b) Seeds growing (c) Distance map

Figure 5. **Visibility modeling without mask annotation. (a)** We first place a set of seeds near the bounding box boundaries. **(b)** We then compute a minimum barrier distance between every pixel within the box and the seeds, **(c)** obtaining a distance map from which we build a probability of visibility for every local object part.

3.2. Visibility-Guided Sampling

The strategy used in the previous experiment relies on the ground-truth object mask during training. However, such masks are typically not available and expensive to obtain, particularly in the presence of occlusions with scene elements. To avoid requiring such masks, we compute an approximate measure of visibility for each pixel in the ground-truth object bounding box.

To this end, let $\mathcal{I} \in \mathbb{R}^{H\times W\times 3}$ be an image patch obtained by cropping a ground-truth object bounding box. We then create a seed set $\mathcal{S} = \{s_1, \cdots, s_m\}$ of 2D positions in the image patch by uniformly sampling the patch boundary with a fixed step size [47, 49]. Our method then builds on the intuition that these seeds will typically *not* belong to the target object. Therefore, the visible object pixels should significantly differ from the seeds, which we encode using an online distance transform.

Specifically, we compute the distance from each pixel within the patch to its nearest seed. For a pixel p and with a generic distance metric, i.e., without assuming the use of the Euclidean distance, this can be expressed as

$$\mathcal{D}(p) = \min_{s\in\mathcal{S}} \mathcal{D}(p, s) \ , \tag{1}$$

where $\mathcal{D}(p, s)$ encodes the distance between pixel p and seed s. Such a distance can in general be expressed as

$$\mathcal{D}(p, s) = \min_{\tau\in\prod_{\{p,s\}}} \mathcal{H}(\tau), \tag{2}$$

where τ is a path connecting pixel p and seed s, $\mathcal{H}(\tau)$ is the cost of path τ, and $\prod_{\{p,s\}}$ is the set containing all possible paths connecting p and s.

Here, we define the cost $\mathcal{H}(\tau)$ as the minimum barrier distance [41], i.e.,

$$\mathcal{H}(\tau) = \mathcal{B}(\mathcal{I}, \tau) + \alpha \cdot d(\tau_0, \tau_1), \tag{3}$$

where τ_0 and τ_1 are the path's starting and ending point, respectively, $d(\tau_0, \tau_1)$ is the Euclidean distance between these two points, and

$$\mathcal{B}(\mathcal{I}, \tau) = \max_{i=1}^{3}(\max_{t=0}^{1} \mathcal{I}_i(\tau_t) - \min_{t=0}^{1} \mathcal{I}_i(\tau_t)), \tag{4}$$

with $\mathcal{I}_i(\tau_t)$ the intensity of the i^{th} channel at a pixel τ_t along the path. We set the balance factor $\alpha = 0.1$ in our experiments, which makes the distance rely mainly on the difference between the maximum and minimum pixel value along the path, thus improving the robustness to different illumination conditions [41].

The resulting distance can be computed efficiently using a fast minimum-barrier-distance solver [18], which lets us generate the corresponding distance transform map $\mathcal{D}(p)$. Fig. 5 illustrates the procedure discussed above, showing that it correctly reflects the object visibility.

In essence, our distance maps provide us with soft visibility masks for the target objects. We then use these soft masks to sample positive cells in a single-stage detection framework, as discussed in Section 3.1.

To this end, for every cell c in every feature map extracted by the FPN module, we compute a visibility score that c belongs to the object as

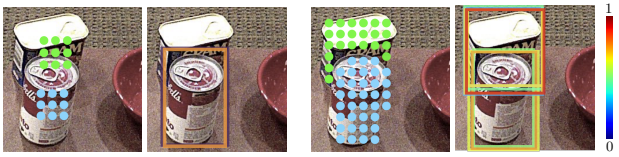
$$\mathcal{V}(c) = \frac{\bar{\mathcal{D}}(c)}{\max_{f\in\mathcal{F}} \bar{\mathcal{D}}(f)}, \tag{5}$$

where $\bar{\mathcal{D}}(c)$ averages the distance map values of all the pixels encompassed by cell c, and $\mathcal{F}$ is the set of all cells in the feature map of interest. We then only consider the cells such that $\mathcal{V}(c) > \mathcal{T}$ as candidate positives, and use $\mathcal{T} = 0.25$ in our experiments.

Note, however, that using all the cells with $\mathcal{V}(c) > \mathcal{T}$ as positives would result in training being dominated by larger objects. To prevent this, we randomly select $k = 10$ cells for each object instance according to $\mathcal{V}(c)$. For the instances containing less than k foreground cells, we randomly sample existing ones multiple times to nonetheless obtain k positive samples. We then discard the cells not chosen as positive samples yet still having a visibility score larger than the threshold $\mathcal{T}$ from the classification and box regression process to avoid providing the network with potentially inconsistent supervision signal.

3.3. Fusion During Inference

As discussed in Section 3.1, during inference, each object instance typically receives multiple box predictions. On the general COCO dataset, Non-Maximum Suppression [43, 44, 50] is typically the method of choice to select a single box, choosing the candidate with the maximum confidence within a local area [1]. This strategy builds on the assumption that only a small region within the box, typically near the box center, can provide a prediction with



(a) The standard strategy (b) The proposed strategy

Figure 6. **Robustness of different strategies.** The left and right parts of **(a)** and **(b)** show the sampling strategy during training and all the valid local predictions before fusion during inference, respectively. The standard center-based sampling strategy suffers from occlusions, as evidenced by the lack of valid predictions for the upper box. Additionally, it generates candidate predictions with large differences in confidence values (as shown by the color difference for the lower box). By contrast, our strategy is robust to occlusions and yields more candidate predictions with high confidence, which can be combined to obtain better results.

high precision, as shown in Fig. 4(a). In the 6D pose estimation setting, however, all visible parts can provide almost equally-accurate predictions, thanks to the rigidity of the targets, as shown in Fig. 4(b).

We therefore propose to combine all the candidate boxes in a neighborhood to obtain a more accurate result. To this end, we let the feature cells predict an additional confidence value, representing how precise the predicted box is. We then cluster the different local predictions that have the same local maximum and assign them to the same object instance. This strategy is similar to the NMS one, but without any candidate suppression. We then compute a simple weighted sum [51] to combine all the candidate local predictions within the same cluster, with weights based on the predicted confidence values. Fig. 6 demonstrates the advantages of this strategy.

3.4. Implementation Details

As mentioned above, we use the same FPN architecture as most state-of-the-art single-stage frameworks [22, 44, 50, 52]. We define the confidence value as the IOU between the predicted box and the ground-truth one. We then train our model with a combined loss function [44, 50]

$$\mathcal{L} = \mathcal{L}_{cls}(\theta, g) + \mathcal{L}_{reg}(\theta, g) + \mathcal{L}_{iou}(\theta, g), \qquad (6)$$

where θ denotes the model parameters and g encodes the ground-truth boxes. $\mathcal{L}_{cls}$ is the focal loss for classification, $\mathcal{L}_{reg}$ is the box regression loss, and $\mathcal{L}_{iou}$ is the cross entropy between the predicted IOU and the ground-truth IOU. We use GIOU [38] loss for $\mathcal{L}_{reg}$ in our implementation.

During training, we first assign every instance to one pyramid level on FPN according to the object size, similarly to [43]. We then compute our distance map on the fly within the annotated bounding box and use it to guide the positive sampling as discussed above. During inference, we use a threshold of 0.05 based on the classification score to remove most of the noise from the background before the clustering and fusing the boxes as discussed in Section 3.3.

4. Experiments

In this section, we systematically study our detection method in 6D object pose estimation scenarios. We first compare its detection performance with other detection baselines in Section 4.1, and then examine its effect when used as bounding box initialization for different pose regression networks in Section 4.2. The code is available at https://github.com/YangHai-1218/RADet.

Experimental settings. We evaluate our method on seven core datasets from BOP [14], including LM-O [2], T-LESS [13], ITODD [8], HB [21], YCB-V [48], IC-BIN [7], and TUD-L [14], which are standard benchmarks for 6D object pose estimation. Most of the datasets have both real images and synthetic ones generated by physically based rendering (PBR) [5] for training, and another split of real images for testing. We use mixed data for training by default. However, for LM-O, ITODD, HB, and IC-BIN, we have only 50k synthetic images for training. As such, we train models only on synthetic images on these datasets.

For a fair comparison with other detection methods, we use the same training setting for both our method and all the competitors unless otherwise stated. We use a ResNet-50 backbone with pre-trained weights from ImageNet [39], a batch size of 16, and an input image resolution fixed at 640×480. We train all the models with the SGD optimizer for 90k iterations, using an initial learning rate of 0.01 with a decay ratio of 0.1 after 60k and 80k iterations, respectively.

Evaluation metrics. We report numbers in the standard metric AP for detection results [28, 43, 50], which is the average value of different AP values obtained with an IOU threshold between the ground truth box and the predicted one ranging from 0.5 to 0.95. For a detailed study, we also report AP_{50} and AP_{75}, which use an IOU threshold of 0.5 and 0.75, respectively.

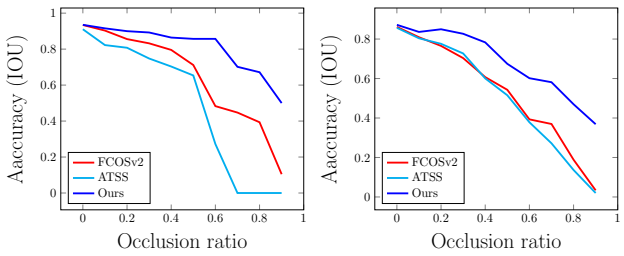
For 6D pose estimation, we report the three standard metrics used in the BOP benchmarks, including the Visible Surface Discrepancy (VSD), the Maximum Symmetry-aware Surface Distance (MSSD), and the Maximum Symmetry-aware Projection Distance (MSPD) [14]. In essence, these metrics differ in the strategies they use to measure the distance between the ground-truth pose and the estimated one. We refer the readers to [14] for their detailed definitions. We report the average numbers of these three metrics in some of our evaluations to save space, and encourage the reader to check our supplementary material for the detailed numbers of each metric.

Method	LM-O	T-LESS	TUD-L	IC-BIN	ITODD	HB	YCB-V	Avg.
Ours	**67.5**	**79.8**	**86.6**	**63.8**	**48.6**	**73.5**	**85.0**	**72.1**
FCOSv2 [44]	57.0	75.0	86.0	27.2	30.4	60.4	80.0	66.7
Mask R-CNN [11]	56.6	69.3	82.6	40.1	36.5	63.5	74.5	60.5

Table 1. **Detection comparison on different 6D object datasets.** Our method achieves much better accuracy than the baseline methods on these BOP datasets, demonstrating the effectiveness of our approach at detecting rigid objects in cluttered 6D pose estimation scenarios.

Method	AP	AP_{50}	AP_{75}
Ours	**85.0**	**99.4**	**97.4**
PAA [22]	83.5	98.3	93.2
AutoAssign [52]	83.3	98.1	91.7
ATSS [50]	82.8	98.0	91.4
FCOSv2 [44]	80.0	98.6	89.1
Faster R-CNN [37]	73.7	92.5	83.3

Table 2. **Detection comparison on YCB-V.** Our method consistently outperforms the state-of-the-art methods, especially in terms of AP_{75}.



(a) YCB-V (b) LM-O

Figure 7. **Performance w.r.t. different occlusion ratios.** Our method is much more robust to occlusions than the baselines.

Method	AP	AP_{50}	AP_{75}
Center†	80.0	98.6	89.1
Center	80.2	98.6	89.5
Ours†	84.2	99.3	96.2
Ours	85.0	99.4	97.4
Oracle†	84.8	**99.6**	97.2
Oracle	**85.7**	**99.6**	**97.9**

Table 3. **Ablation study on YCB-V.** We compare the center-based and the proposed visibility-guided sampling strategies used with standard NMS (denoted by †) or with our fusion strategy. Our method is more accurate than the baseline strategies and performs on par with the oracle one that uses on guided sampling from the ground-truth mask. Here, "Center†" corresponds to the strategy of FCOSv2 [44].

4.1. Object Detection

Comparison with the baselines. We compare our method with the baseline single-stage method FCOSv2 [44] and a typical two-stage method, Mask R-CNN [11]. As shown in Table 1, our method outperforms them by a large margin on all datasets from the BOP benchmarks, demonstrating the effectiveness of our approach at detecting rigid objects in cluttered 6D pose estimation scenarios.

Comparison with the state of the art on YCB-V. We compare our method with the state-of-the-art detection methods, including AutoAssign [52] and PAA [22], on the YCB-V dataset. Table 2 summarizes the results, showing that our method consistently outperforms the state-of-the-art ones, especially in terms of AP_{75}.

Performance under occlusions. Occlusion is a common problem in most BOP benchmarks. We study the impact of different occlusion levels on different detectors. We compare our method with FCOSv2 [44] and ATSS [50] on both YCB-V and LMO, and compute the average accuracy of the results with respect to the targets' occlusion ratio. The results are as summarized in Fig. 7. Although ATSS improves the center-based sampling of FCOS by its adaptive assignment strategy across multiple pyramid levels, it remains sensitive to occlusions, as illustrated by the quick deterioration of accuracy with the increasing of occlusion ratio. By contrast, our method is much more robust.

Ablation study on YCB-V. We compare NMS and our fusion strategy on a model trained either with the centered-based strategy or the proposed visibility-guided one. As shown in Table 3, with the same NMS post-processing, our sampling strategy already outperforms the center-based baseline by 4.2 points. This confirms the importance of involving all the visible object parts during training, leveraging the rigidity of the targets. Furthermore, both sampling strategies benefit from our fusion method discussed in Section 3.3. However, it only increases the performance of the center-based one by 0.2 points, which highlights the drawback of not using non-center areas during training, making the fusion during inference less effective. By contrast, our fusion method increases the performance of our sampling strategy by 0.8 points, making it perform on par with the oracle that uses the ground-truth mask to guide sampling.

Runtime analysis. We conduct all our experiments on a

Method	Real	Data	LM-O*	T-LESS	TUD-L	IC-BIN*	ITODD*	HB*	YCB-V	Avg.
PFA+Ours		RGB	**0.715**	0.719	**0.733**	**0.600**	0.353	**0.840**	**0.648**	**0.658**
PFA [15]		RGB	0.674	-	-	-	-	-	0.614	-
SurfEmb [10]		RGB	0.663	**0.735**	0.715	0.588	**0.413**	0.791	0.647	0.650
CosyPose [23]		RGB	0.633	0.640	0.685	0.473	0.216	0.656	0.574	0.570
CDPNv2 [26]		RGB	0.624	0.407	0.588	0.226	0.067	0.722	0.390	0.472
PFA+Ours	✓	RGB	**0.715**	**0.778**	**0.839**	**0.600**	0.353	**0.840**	0.806	**0.704**
PFA [15]	✓	RGB	0.674	-	-	-	-	-	0.748	-
SurfEmb [10]	✓	RGB	0.663	0.770	0.805	0.588	**0.413**	0.791	0.711	0.677
CosyPose [23]	✓	RGB	0.633	0.728	0.823	0.583	0.216	0.656	**0.821**	0.637
CDPNv2 [26]	✓	RGB	0.624	0.478	0.772	0.473	0.067	0.722	0.532	0.529
PFA+Ours		RGBD	**0.797**	0.801	**0.894**	**0.676**	0.460	**0.869**	**0.826**	**0.762**
PFA [15]		RGBD	0.751	-	-	-	-	-	0.804	-
SurfEmb [10]		RGBD	0.760	**0.828**	0.854	0.659	**0.538**	0.866	0.799	0.758
CDPNv2+ICP [26]		RGBD	0.630	0.435	0.791	0.450	0.186	0.712	0.532	0.534
PFA+Ours	✓	RGBD	**0.797**	**0.850**	0.960	**0.676**	0.460	**0.869**	0.888	**0.787**
PFA [15]	✓	RGBD	0.751	-	-	-	-	-	0.823	-
SurfEmb [10]	✓	RGBD	0.760	0.833	0.933	0.659	**0.538**	0.866	0.824	0.773
CIR [30]	✓	RGBD	0.734	0.776	**0.968**	**0.676**	0.381	0.757	**0.893**	0.741
CosyPose+ICP [23]	✓	RGBD	0.714	0.701	0.939	0.647	0.313	0.712	0.861	0.698
CDPNv2+ICP [26]	✓	RGBD	0.630	0.464	0.913	0.450	0.186	0.712	0.619	0.568

Table 4. **Comparison against the state of the art on 6D pose estimation.** Our detection method improves the original PFA-Pose by a large margin, and yields state-of-the-art results with either only synthetic or mixed data in both the RGB and RGBD settings. Note that LM-O [2], IC-BIN [7], ITODD [8], and HB [21] provide only the synthetic PBR images for training, so the numbers "w/o Real" and "w/ Real" are the same on those datasets, indicated by "$*$". Here, we report the results as the average of the three metrics, MSPD, MSSD, and VSD.

Method	Avg.	MSPD	MSSD	VSD
WDR + **Ours**	**0.605**	**0.694**	**0.598**	**0.522**
WDR + RCNN	0.587	0.673	0.580	0.508
WDR + FCOSv2	0.585	0.671	0.578	0.506
CDPNv2 + **Ours**	**0.412**	**0.534**	**0.428**	**0.275**
CDPNv2 + FCOSv2	0.402	0.523	0.416	0.268
CDPNv2 + RCNN	0.388	0.506	0.401	0.258

Table 5. **Effect on different pose regression networks.** Our detection method consistently improves the results of different pose regression frameworks, including WDR [19] and CDPNv2 [26]. Here, we report numbers on YCB-V, and denote Mask R-CNN [11] as "RCNN'.

workstation with an NVIDIA RTX-3090 GPU and an Intel-Xeon CPU with 12 2.1GHz cores. Our method shares the same network architecture as most single-stage methods [22, 44, 50, 52] and the running time of our simple fusion strategy is negligible. As such, all methods have a similar inference speed of about 32.4 images per second on the YCB-V dataset with an average of 4.8 instances in each image. The main difference comes from the training time, since different methods rely on different sampling strategies. Our method has a throughput of about 18.7 images per second during training, which is slightly slower than FCOSv2 (22.3) and ATSS (21.6), but faster than PAA (16.6) and AutoAssign (15.7).

4.2. Object Pose Estimation

Comparison with the state of the art. To demonstrate the effectiveness of our detection method in 6D object pose estimation, we combine it with a recent pose regression network, PFA-Pose [15], and compare the pose results with other methods. We test our method with PFA-Pose in different settings, including training only on synthetic PBR or with mixed real images. Additionally, we evaluate our method when PFA-Pose uses a simple depth refinement strategy based on RANSAC-Kabsch [20,40] to consume additional depth images. The original PFA-Pose cannot handle multiple instances from the same class, making it inapplicable to some datasets. So we only reproduce its results on LM-O and YCB-V. Table 4 summarizes the results, showing that our detection method improves the original PFA-Pose by a large margin, obtaining state-of-the-art pose estimation results with either only synthetic or mixed data

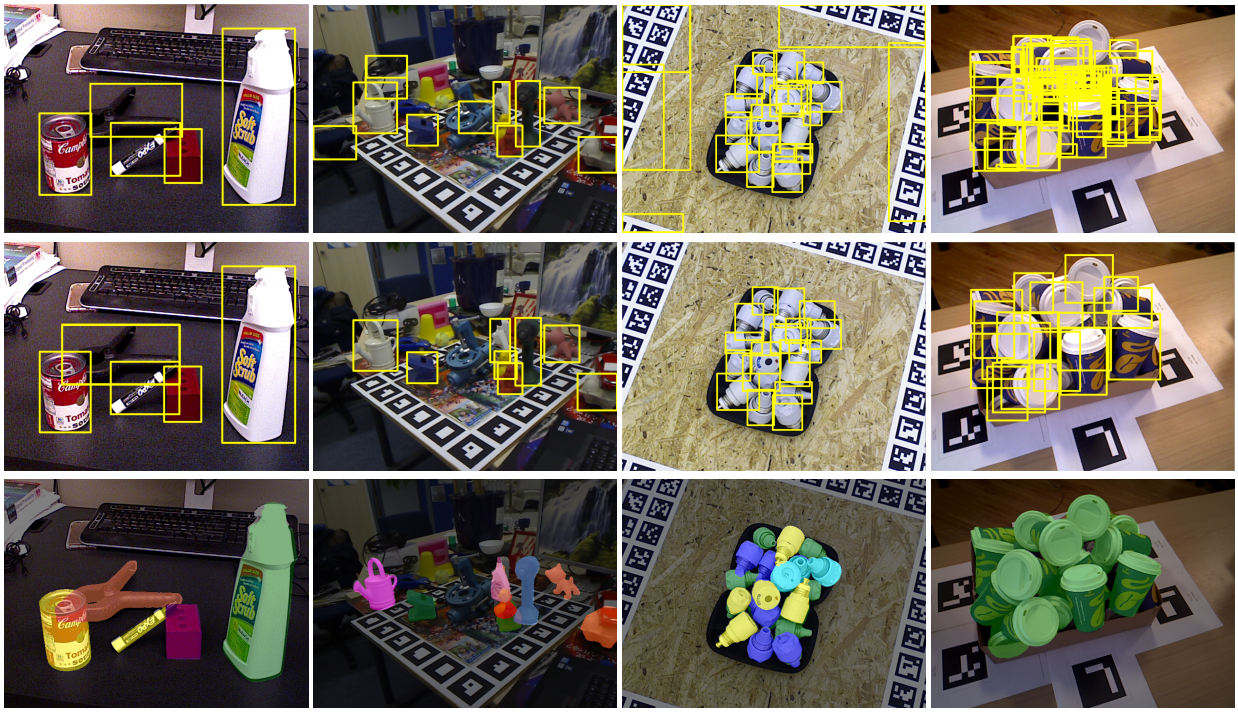


Figure 8. **Visualization of detection and pose results.** The first and second rows show the detection results of the baseline FCOS [43] and our method on different datasets (LM-O, YCB-V, T-LESS, and IC-BIN), respectively. Although the baseline works almost equally well in simple cases, such as targets without occlusions, it deteriorates significantly for targets in cluttered scenes, and generate many more false positives. By contrast, our detection method is robust, and produces accurate pose estimates after using a subsequent pose regression network (PFA [15]), as shown in the last row.

in both the RGB and RGBD settings. Fig. 8 visualizes some results.

Evaluation with different pose regression networks. In principle, our detection method can be used with most pose regression frameworks as a first component to extract the object's bounding box before pose regression. To demonstrate its generalization ability, we test our detection method with two other typical pose regression networks, WDR-Pose [19] and CDPNv2 [26]. Table 5 provides the results, evidencing that our detection method consistently improves the pose results.

5. Conclusion

We have proposed a visibility-guided sampling strategy for training a deep network to detect rigid objects in cluttered scenes.We have first analyzed the influence of the rigidity of the targets in the 6D object pose estimation scenarios and studied the weaknesses of general detection methods in this setting. Based on the observation that detecting rigid objects should allow us to rely on all visible object parts, and that each part should already provide a reliable prediction of the whole bounding box, we have proposed to build a visibility map to guide the positive sampling during training and combine multiple local predictions during inference to obtain the final robust result. We have demonstrated the effectiveness of our method on the challenging datasets from the BOP benchmarks. It achieves much better detection results than general methods, and produces state-of-the-art pose results when combined with pose regression networks.

Figure 9. **Examples of the difficulty in approximating visibility.**

Discussion. Although our visibility-guided sampling performs on par with the oracle one that uses the ground-truth mask to guide sampling, it still has some gaps from the oracle one, which is mainly caused by the difficulty of establishing reliable distance map in clutter scenarios with severe occlusions and complex textures, as shown in Fig. 9. We will investigate strategies for learning the visibility of targets from data in the future.

Acknowledgments. This work was supported in part by the National Nature Science Foundation of China under Grant 61901343, Grant 61801359, Grant 61701360, Grant 61671383, and Grant 61571345; in part by the 111 Project (B08038); in part by the Fundamental Research Funds for the Central Universities; in part by The Youth Innovation Team of Shaanxi Universities, and in part by the Wuhu and Xidian University special fund for industry-university-research cooperation(No.XWYCXY-012021002). We would like to thank Zhaoyang Liu and Wayne Wu for the helpful discussions in the project's early stage.

References

[1] Navaneeth Bodla, Bharat Singh, Rama Chellappa, and Larry S. Davis. Soft-nms - improving object detection with one line of code. In *International Conference on Computer Vision*, 2017. 4

[2] Eric Brachmann, Alexander Krull, Frank Michel, Stefan Gumhold, Jamie Shotton, and Carsten Rother. Learning 6d object pose estimation using 3d object coordinates. In *European Conference on Computer Vision*, 2014. 5, 7

[3] Dingding Cai, Janne Heikkilä, and Esa Rahtu. Sc6d: Symmetry-agnostic and correspondence-free 6d object pose estimation. In *International Conference on 3D Vision*, 2022. 1, 2

[4] Zhaowei Cai and Nuno Vasconcelos. Cascade R-CNN: delving into high quality object detection. In *Conference on Computer Vision and Pattern Recognition*, 2018. 2

[5] Maximilian Denninger, Martin Sundermeyer, Dominik Winkelbauer, Dmitry Olefir, Tomas Hodan, Youssef Zidan, Mohamad Elbadrawy, Markus Knauer, Harinandan Katam, and Ahsan Lodhi. BlenderProc: reducing the reality gap with photorealistic rendering. *Robotics: Science and Systems Workshops*, 2020. 5

[6] Yan Di, Fabian Manhardt, Gu Wang, Xiangyang Ji, Nassir Navab, and Federico Tombari. So-pose: Exploiting self-occlusion for direct 6d pose estimation. In *International Conference on Computer Vision*, 2021. 1, 2

[7] Andreas Doumanoglou, Rigas Kouskouridas, Sotiris Malassiotis, and Tae-Kyun Kim. Recovering 6d object pose and predicting next-best-view in the crowd. In *Conference on Computer Vision and Pattern Recognition*, 2016. 5, 7

[8] Bertram Drost, Markus Ulrich, Paul Bergmann, Philipp Hartinger, and Carsten Steger. Introducing mvtec itodd-a dataset for 3d object recognition in industry. In *International Conference on Computer Vision Workshops*, 2017. 5, 7

[9] Zheng Ge, Songtao Liu, Zeming Li, Osamu Yoshie, and Jian Sun. Ota: Optimal transport assignment for object detection. In *Conference on Computer Vision and Pattern Recognition*, 2021. 2

[10] Rasmus Laurvig Haugaard and Anders Glent Buch. Surfemb: Dense and continuous correspondence distributions for object pose estimation with learnt surface embeddings. In *Conference on Computer Vision and Pattern Recognition*, 2022. 1, 2, 7

[11] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B. Girshick. Mask R-CNN. In *International Conference on Computer Vision*, 2017. 1, 2, 6, 7

[12] Tomáš Hodaň, Dániel Baráth, and Jiří Matas. EPOS: Estimating 6D pose of objects with symmetries. In *Conference on Computer Vision and Pattern Recognition*, 2020. 2

[13] Tomáš Hodan, Pavel Haluza, Štepán Obdržálek, Jiri Matas, Manolis Lourakis, and Xenophon Zabulis. T-less: An rgb-d dataset for 6d pose estimation of texture-less objects. In *Winter Conference on Applications of Computer Vision*, 2017. 5

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

[15] Yinlin Hu, Pascal Fua, and Mathieu Salzmann. Perspective flow aggregation for data-limited 6d object pose estimation. In *European Conference on Computer Vision*, 2022. 1, 2, 7, 8

[16] Yinlin Hu, Pascal Fua, Wei Wang, and Mathieu Salzmann. Single-stage 6d object pose estimation. In *Conference on Computer Vision and Pattern Recognition*, 2020. 2

[17] Yinlin Hu, Joachim Hugonot, Pascal Fua, and Mathieu Salzmann. Segmentation-driven 6d object pose estimation. In *Conference on Computer Vision and Pattern Recognition*, 2019. 2

[18] Yinlin Hu, Yunsong Li, Rui Song, Peng Rao, and Yangli Wang. Minimum barrier superpixel segmentation. *Image and Vision Computing*, 70:1 – 10, 2018. 4

[19] Yinlin Hu, Sebastien Speierer, Wenzel Jakob, Pascal Fua, and Mathieu Salzmann. Wide-depth-range 6d object pose estimation in space. In *Conference on Computer Vision and Pattern Recognition*, 2021. 2, 7, 8

[20] Lin Huang, Tomas Hodan, Lingni Ma, Linguang Zhang, Luan Tran, Christopher Twigg, Po-Chen Wu, Junsong Yuan, Cem Keskin, and Robert Wang. Neural correspondence field for object pose estimation. In *European Conference on Computer Vision*, 2022. 2, 7

[21] Roman Kaskman, Sergey Zakharov, Ivan Shugurov, and Slobodan Ilic. Homebreweddb: Rgb-d dataset for 6d pose estimation of 3d objects. In *International Conference on Computer Vision Workshops*, 2019. 5, 7

[22] Kang Kim and Hee Seok Lee. Probabilistic anchor assignment with iou prediction for object detection. In *European Conference on Computer Vision*, 2020. 1, 2, 3, 5, 6, 7

[23] Yann Labbé, Justin Carpentier, Mathieu Aubry, and Josef Sivic. Cosypose: Consistent multi-view multi-object 6d pose estimation. In *European Conference on Computer Vision*, 2020. 2, 7

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

[25] Yi Li, Gu Wang, Xiangyang Ji, Yu Xiang, and Dieter Fox. Deepim: Deep iterative matching for 6d pose estimation. In *European Conference on Computer Vision*, 2018. 1

[26] Zhigang Li, Gu Wang, and Xiangyang Ji. Cdpn: Coordinates-based disentangled pose network for real-time rgb-based 6-dof object pose estimation. In *International Conference on Computer Vision*, 2019. 2, 7, 8

[27] Tsung-Yi Lin, Piotr Dollár, Ross B. Girshick, Kaiming He, Bharath Hariharan, and Serge J. Belongie. Feature pyramid networks for object detection. In *Conference on Computer Vision and Pattern Recognition*, 2017. 2, 3

[28] Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *International Conference on Computer Vision*, 2017. 2, 3, 5

[29] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. In *European Conference on Computer Vision*, 2014. 1, 3

[30] Lahav Lipson, Zachary Teed, Ankit Goyal, and Jia Deng. Coupled iterative refinement for 6d multi-object pose estimation. In *Conference on Computer Vision and Pattern Recognition*, 2022. 7

[31] Zili Liu, Tu Zheng, Guodong Xu, Zheng Yang, Haifeng Liu, and Deng Cai. Training-time-friendly network for real-time object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020. 2

[32] Kiru Park, Timothy Patten, and Markus Vincze. Pix2pose: Pixel-wise coordinate regression of objects for 6d pose estimation. In *International Conference on Computer Vision*, 2019. 2

[33] Sida Peng, Yuan Liu, Qixing Huang, Xiaowei Zhou, and Hujun Bao. Pvnet: Pixel-wise voting network for 6dof pose estimation. In *Conference on Computer Vision and Pattern Recognition*, 2019. 2

[34] Joseph Redmon, Santosh Kumar Divvala, Ross B. Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Conference on Computer Vision and Pattern Recognition*, 2016. 2

[35] Joseph Redmon and Ali Farhadi. YOLO9000: better, faster, stronger. In *Conference on Computer Vision and Pattern Recognition*, 2017. 2

[36] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *CoRR*, abs/1804.02767, 2018. 2

[37] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, 2015. 1, 2, 6

[38] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian D. Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *Conference on Computer Vision and Pattern Recognition*, 2019. 5

[39] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 2015. 5

[40] Ivan Shugurov, Sergey Zakharov, and Slobodan Ilic. Dpodv2: Dense correspondence-based 6 dof pose estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11):7417–7435, 2021. 2, 7

[41] Robin Strand, Krzysztof Chris Ciesielski, Filip Malmberg, and Punam K Saha. The minimum barrier distance. *Computer Vision and Image Understanding*, 117(4):429–437, 2013. 4

[42] Yongzhi Su, Mahdi Saleh, Torben Fetzer, Jason Rambach, Nassir Navab, Benjamin Busam, Didier Stricker, and Federico Tombari. Zebrapose: Coarse to fine surface encoding for 6dof object pose estimation. In *Conference on Computer Vision and Pattern Recognition*, 2022. 1, 2

[43] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. FCOS: fully convolutional one-stage object detection. In *International Conference on Computer Vision*, 2019. 1, 2, 3, 4, 5, 8

[44] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: A simple and strong anchor-free object detector. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(4):1922–1933, 2020. 1, 2, 3, 4, 5, 6, 7

[45] Chen Wang, Danfei Xu, Yuke Zhu, Roberto Martin Martin, Cewu Lu, Li Fei-Fei, and Silvio Savarese. Densefusion: 6d object pose estimation by iterative dense fusion. In *Conference on Computer Vision and Pattern Recognition*, 2019. 1

[46] Gu Wang, Fabian Manhardt, Federico Tombari, and Xiangyang Ji. Gdr-net: Geometry-guided direct regression network for monocular 6d object pose estimation. In *Conference on Computer Vision and Pattern Recognition*, 2021. 2

[47] Yichen Wei, Fang Wen, Wangjiang Zhu, and Jian Sun. Geodesic saliency using background priors. In *European Conference on Computer Vision*, 2012. 4

[48] 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*, 2018. 3, 5

[49] Jianming Zhang, Stan Sclaroff, Zhe Lin, Xiaohui Shen, Brian Price, and Radomir Mech. Minimum barrier salient object detection at 80 fps. In *International Conference on Computer Vision*, 2015. 4

[50] Shifeng Zhang, Cheng Chi, Yongqiang Yao, Zhen Lei, and Stan Z. Li. Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection. In *Conference on Computer Vision and Pattern Recognition*, 2020. 1, 2, 3, 4, 5, 6, 7

[51] Huajun Zhou, Zechao Li, Chengcheng Ning, and Jinhui Tang. Cad: Scale invariant framework for real-time object detection. In *International Conference on Computer Vision Workshops*, 2017. 5

[52] Benjin Zhu, Jianfeng Wang, Zhengkai Jiang, Fuhang Zong, Songtao Liu, Zeming Li, and Jian Sun. Autoassign: Differentiable label assignment for dense object detection, 2020. 2, 3, 5, 6, 7