

Multi-Task Deep Networks for Depth-Based 6D Object Pose and Joint Registration in Crowd Scenarios

Juil Sock¹
ju-il.sock08@imperial.ac.uk
Kwang In Kim²
k.kim@bath.ac.uk
Caner Sahin¹
c.sahin14@imperial.ac.uk
Tae-Kyun Kim¹
tk.kim@imperial.ac.uk

¹ Imperial Computer Vision and Learning Lab (ICVL), Imperial College, London, UK
² Department of Computer Science, University of Bath, Bath, UK

Abstract

In bin-picking scenarios, multiple instances of an object of interest are stacked in a pile *randomly*, and hence, the instances are inherently subjected to the challenges: *severe occlusion*, *clutter*, and *similar-looking distractors*. Most existing methods are, however, for single isolated object instances, while some recent methods tackle crowd scenarios as post-refinement which accounts for multiple object relations. In this paper, we address recovering 6D poses of multiple instances in bin-picking scenarios in depth modality by multi-task learning in deep neural networks. Our architecture *jointly* learns multiple sub-tasks: 2D detection, depth, and 3D pose estimation of individual objects; and joint registration of multiple objects. For training data generation, depth images of physically plausible object pose configurations are generated by a 3D object model in a physics simulation, which yields diverse occlusion patterns to learn. We adopt a state-of-the-art object detector to generate bounding boxes and 2D offsets are further estimated via a network to refine misalignments in 2D detections. Additionally, the depth and 3D pose estimators are designed to generate multiple hypotheses per detection. This allows the joint registration network to learn occlusion patterns and remove physically implausible pose hypotheses. We apply our architecture on both synthetic (our own and Siléane dataset) and real (a public Bin-Picking dataset) data, showing that it significantly outperforms state-of-the-art methods by 15-31 % in average precision.

1 Introduction

6D object pose recovery has been extensively studied in the past decade as it is of great importance to many real-world applications, such as scene interpretation, augmented reality, and particularly, robotic grasping and manipulation. Robotic manipulators can facilitate industry-related processes, *e.g.*, warehousing, material handling, packaging, picking items from conveyors, shelves, pallets, and placing those to desired locations, as demonstrated in

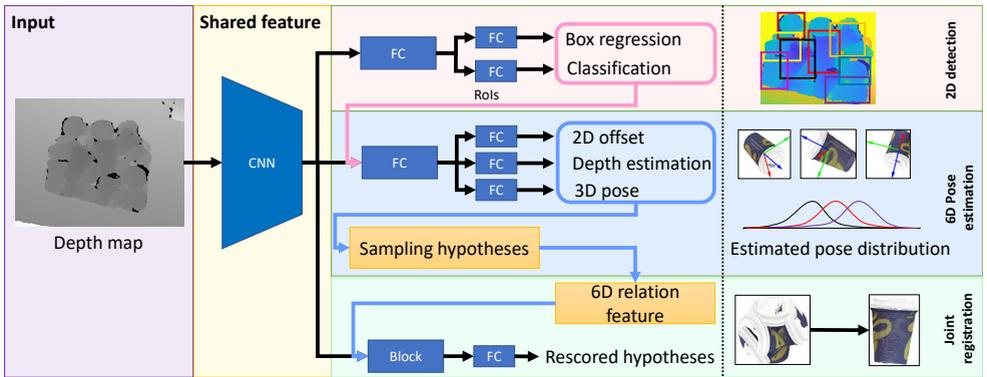


Figure 1: Our multi-task learning network architecture consists of three modules, each optimized for different sub-tasks: 2D detection, 6D pose estimation, and joint registration. During inference, each sub-tasks are sequentially performed (red and blue arrows). During training, the network is trained jointly using multi-task losses, generating the convolutional features shared across sub-tasks.

Amazon Picking Challenge (APC) [9].

A special scenario for robotic manipulation is bin-picking, where multiple instances of an object of interest exist in the scene. Being randomly stacked in a box, or on a pallet, the instances are inherently subjected to severe occlusion and clutter, thus making the estimation of the 6D poses of multiple instances challenging.

Most existing methods considers single object instance [10] scenario, while few recent methods tackle crowd scenarios in post-refinement which accounts for multiple object relations [1]. Prior-arts for 6D object pose include template-based methods [11], point-pair features [8], or per-pixel regression/patch-based on top of Hough forests [4, 7, 12, 17]. These methods are mostly applied and tested on single object scenario with mild occlusion. More recent methods adopt CNNs for 6D pose estimation, taking RGB images as inputs. BB8 [17] and Tekin *et al.* [19] perform corner-point regression followed by PnP for 6D pose estimation. SSD6D [16] uses SSD for 2D bounding box detection, while depth is directly inferred from the scales of bounding boxes, and applies a network with a softmax layer for 3D pose. The method has shown comparable performance to other RGBD-based methods, however, fails in the occlusion dataset [9].

For the crowd scenarios, *i.e.*, bin-picking, rather than considering each hypothesis individually, [1] formulates a joint registration problem. It obtains an optimal set of hypotheses of multiple objects by minimizing the cost function. Doumanoglou *et al.* [1] collect a real dataset of bin-picking scenarios, proposing 6D pose estimator based on a sparse auto-encoder and hough forest, and then tackling the joint registration in the similar way [1]. Recently, [5] proposes a dataset of both synthetic and real in bin-picking scenarios, and [20] tackles deformable objects (fruits) in crowd by matching local descriptors.

In the domain of 2D object detection, recent methods have shown that learning the relation between nearby hypotheses in crowd scenario can increase the detection performance by effectively suppressing false hypotheses. [13] proposes a method to learn non maximum suppression for 2D bounding boxes. More recently, [14] trains a 2D detection network jointly with relation module in an end-to-end manner.

In this paper, we propose a depth-based 6D pose estimation network which jointly detects and directly recovers (without ICP) full 6D poses from single depth images. Depth image is

used in the proposed architecture, as instances are densely populated in crowd scenario and geometric information from depth image is desirable. Directly estimating multiple 6D poses of objects from a single image is a difficult task, therefore the architecture has modules optimized for different sub-tasks: image 2D detection, depth estimation and 3D pose estimation for individual objects, and joint registration of multiple objects as shown in Fig. 1. Although the full architecture consists of several modules, the tasks are jointly learned. The proposed network explicitly learns physically plausible object pose configurations and occlusion patterns from depth images generated by physics simulator and 3D models of objects.

2D offset estimation regresses 2D center position of the object given the detected box coordinate to compensate misalignment during the bounding box regression. Multiple hypotheses per detected object are generated by the depth and 3D pose estimator by formulating the estimation problem as a classification task. Lastly, relations between hypotheses from the pose estimator are modeled by the joint registration network using block modules [13] to suppress false positives. Our architecture is evaluated on both synthetic and real datasets demonstrating that it significantly outperforms state-of-the-art methods by 15-31 % in AP.

2 Related Work

In this section, we first review 6D object pose recovery for *single object instances*, and then for *crowd scenarios*, i.e., bin-picking. Lastly, we present studies on *relational reasoning*, which is used to further improve the confidence of a 6D hypothesis in our study.

Single object instances. Under moderate occlusion and clutter, accurate pose estimation has been achieved by approaches with handcrafted features [8, 10, 27]. 3D feature proposed in [8] can deal with clutter and occlusion but the performance depends on the complexity of the scene and it is computationally expensive. A holistic template matching approach, Linemod [10], exploits color gradients and surface normals to estimate 6D pose. However, it is sensitive to occlusion as it is difficult to match occluded object to template captured from a clean view of the object. More recently, per-pixel regression/patch-based approaches [9, 11, 17, 27] has shown robustness across occlusion and clutter. [9, 17] use random forest to obtain pixel-wise prediction of class and object coordinate followed by RANSAC. However, performance was only evaluated on scenes with mild occlusion and limited variability of object poses relative to the camera. LCHF[27] and Domanoglou *et al.* [11] use patch-based approach on top of hough voting and has shown strong performance under moderate occlusion. However, detection of objects under strong occlusion is difficult as the occluded object only receives vote from visible portion in hough space.

More recent paradigm on recovering 6D object pose is to formulate the problem using neural networks[3], jointly learning 6D pose estimation in RGB-only images [16, 29, 31]. [16] extends 2D object detector to simultaneously detect and estimate pose and recover 3D translation by precomputing bounding box templates for every discrete rotation. However, predicted bounding boxes can be highly variable in the presence of occlusion which may lead to incorrect 3D translation estimation. Rad *et al.* [29] regress 3D bounding box coordinates in 2D image plane and utilizes PnP algorithm to recover the full 6D pose. However, the method uses segmentation masks to detect objects and thus making the hypotheses unreliable when objects of the same class are in high proximity. Despite being robust across occlusion and clutter under non-crowd scenarios, the methods [16, 29] require a final refinement step in order to generate accurate hypotheses. [29] does not need post-processing. However, the method aggregates estimation from multiple cells under the assumption that adjacent cells in

the tessellated input image belong to the same object, limiting the scalability of the method to isolated scenarios.

Crowd scenarios. Bin-picking scenario is a very challenging task for any vision algorithm since it exhibits several cascading issues (*e.g.*, severe occlusions, foreground clutter, etc.). The method proposed by Liu *et al.* [19] extracts robust depth edges and employs fast directional chamfer matching to reliably detect objects and estimate their poses. [20] obtains an optimal set of hypotheses of multiple objects by minimizing a cost function, rather than considering each hypothesis individually. [7] proposes 6D pose estimator of individual objects based on a sparse auto-encoder and hough forest tackling the joint registration in a similar way as [20]. However, the method is computationally demanding with increasing number of hypotheses. [21] tackles deformable objects (fruits) in the crowd by matching local descriptors. Recently, Zeng *et al.* [32] address the challenges of the crowd scenarios proposing a self-supervised deep learning method for 6D pose estimation in Amazon Picking Challenge.

Relational reasoning. Although most modern object detectors generate detection hypotheses independently, recent works have shown exploiting the relation between hypotheses helps under high occlusion. [18] models the relation between detection hypotheses and uses determinantal point process to select the most confident hypothesis. [26] proposes a network for person detection which eliminates the non maximum suppression stage by modeling the relation between detections using LSTM [22]. Gossipnet[23] performs non-maximum suppression by proposing a block module which develops a representation for each detection which accounts for other nearby detections. [24] adds an extra layer called relation module to the standard object detection framework and jointly trains in an end-to-end fashion. It has shown that learning object detector with relational reasoning improves detection performance. Both methods [23, 24] use relational reasoning to remove duplicate hypotheses in 2D detection. In this work, the architecture developed in [23] is extended to 6D pose estimation problem.

3 Proposed Architecture

A key challenge in object pose estimation in bin-packing scenarios is that multiple object instances exhibit severe occlusions under environments cluttered with the same object. For severely occluded objects, the 6D pose predictions made based on visible parts are inherently ambiguous (see Fig. 2(b)). Therefore, a joint registration step exploiting the neighborhood context (of how the pose of an object is related to neighboring objects) is essential to making reliable predictions. Existing approaches address this issue only at the testing stage, *e.g.*, using non-maximum suppression (NMS) and the training of their pose estimators is agnostic to such occlusion behaviors: They are trained on isolated object instances. Instead, we directly incorporate the variety of occlusion patterns and the possibility of exploiting the neighborhood context into the *training* phase. To facilitate this process, we construct a new training dataset that reflects such a realistic occlusion-heavy application environment.

Figure 1 shows our pose estimator architecture consisting of three modules, each instantiated as a deep neural network: The detection module (red) estimates 2D (along the image plane in the depth map) object locations as bounding boxes. Within each hypothesized bounding box, 6D pose estimation module (blue) estimates the center of the object (which may differ from the center of its bounding box) and its depth (completing the 3D locations), and based on them, generates multiple 3D pose hypotheses (as singling out a correct pose

is challenging under heavy occlusions). Finally, the joint registration module (green) disambiguate between multiple 6D pose hypotheses and eliminate false positives by considering local neighborhood structures. Exploiting the dependency present in these three tasks and the corresponding modules, we formulate their joint training as multi-task learning (MTL): All three modules share the same depth feature extractor (yellow block in Fig. 1), enabling the training of convolutional features in a unified manner.¹

The overall loss function (of network weights) is defined per depth image as:

$$L = L_D + L_O + L_D + L_P, \quad (1)$$

with L_D , L_O , L_D , and L_P being the losses for estimating 2D bounding boxes, 2D object centers, 3D depths, and 6D poses, respectively. During training, the backward propagation takes place as usual for each module, and they are subsequently combined to update the shared feature extractor in Fig. 1. The 2D detection and 6D pose estimation modules are simultaneously updated while for the update of the joint registration module, we hold the rest of the network weights fixed for improved convergence. The remainder of this section details each module and the corresponding individual loss function in Eq. 1.

Generating training images. Unlike most existing 6D pose estimators which learn object appearance from *isolated views* of objects, our system jointly learns object appearances and their mutual occlusion patterns based on cluttered scene images. Figure 2(a) illustrates the generation of training data in this scenario: First, we construct a cluttered scene in a virtual environment by randomly stacking the object of interest in a box similarly to [8, 23] (Fig. 2(a) left). Each object has 2,000 different bin picking scenes, each of which has 10 to 20 instances in the virtual bin. Then, we render training depth images by sampling virtual camera views (Fig. 2(a) middle): The camera is always facing towards the center of the bin but with randomized in-plane rotations to capture various occlusion patterns. 17 depth images are generated from each scene constituting a set of 34K images per object.

2D bounding box detection. In a bin-picking scenario, the number of objects in a given scene is not known a priori. Therefore, the traditional detection strategy of predicting the single most probable location is not applicable. We employ Faster RCNN [22] which has shown its effectiveness in detecting multiple objects in cluttered scenes. While Faster RCNN was originally developed for RGB images, applying it to depth maps is straightforward. The network receives a single image and generates a bounding box and its confidence per object instance: During training, it minimizes the box regression error (via L_1 -loss), and for the detection confidence, the soft classification error (via cross entropy) of foreground and background. We denote the combination of the two losses as L_D (see Eq. 1).

6D pose estimation. The 6D pose of an object consists of its 3D location and 3D pose (orientation). Given the bounding box of each object instance estimated by 2D detection module, 6D pose estimation module generates multiple hypothesized 6D poses, which are subsequently fed to the joint registration module.

2D object center estimation. For each bounding box, we first estimate the accurate 2D location of the corresponding object center. In existing 6D pose estimation approaches (developed for uncluttered scenes) [8, 17], the center of an estimated box is considered identical to the center of the object it contains. For uncluttered scenes, this is reasonable as often, the

¹As there are non-differentiable paths joining the three modules (along the red and blue arrows in Fig. 1), the training is not end-to-end.

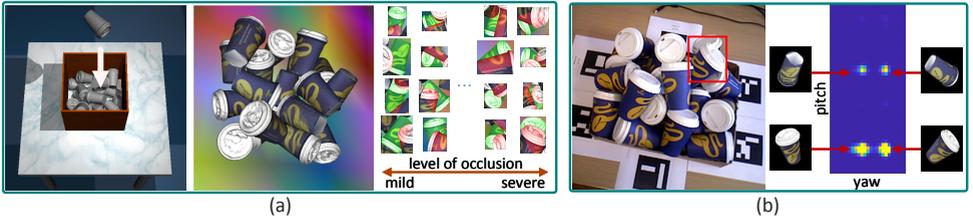


Figure 2: **(a)** Training data generation process: (left) A cluttered scene generated via physical simulation; (middle) An example training view rendered from the scene in (left); (right) Example occlusion patterns (for visualization only; our network is trained on holistic images). **(b)** (left) An example detection under severe occlusion (overlaid on the RGB image for visualization); (right) Multiple hypotheses with confidences (the brighter is the higher in the heatmap) generated by the 6D pose estimator.

estimated boxes are well-aligned with the underlying tight ground-truth bounding boxes. However, we experimentally observed that for objects under severe occlusion, such accurate bounding box estimation is extremely challenging, and naively basing the joint registration on such erroneous 2D location estimates eventually leads to erroneous 6D pose predictions (see Table 1(a) ‘Ours (w/o offset)’). Therefore, we construct an additional 2D regression network that explicitly estimate the object centers by minimizing the deviation from the ground-truth within each bounding box:

$$L_O = \left\| \left[\frac{x - c_x}{w}, \frac{y - c_y}{h} \right]^T - \left[\frac{x^* - c_x}{w}, \frac{y^* - c_y}{h} \right]^T \right\|_1, \quad (2)$$

where (x, y) and (x^*, y^*) denote the predicted and the ground-truth object center coordinates, respectively, (w, h) represents the width and height of the detected box, and (c_x, c_y) is the top-left corner coordinate of the box.

3D pose and depth estimation. A common approach to 3D pose estimation is regression, *e.g.*, to estimate the quaternion [20] or the elements of rotation matrices [9]. However, in the preliminary experiments, we observed that when the object undergoes severe occlusion, generating a single reliable pose estimate is challenging: Inspected at only small portions of objects, indeed, significantly different pose configurations can be instantiated from visually very similar depth values (see Fig. 2(b) for example). To circumvent this challenge, we first generate multiple pose hypotheses and disambiguate them afterward by exploiting the context of neighboring detections and their pose hypotheses during joint registration. To facilitate hypothesis generation, we adopt the classification approach of Kehl *et al.* [16]: Three rotation components are estimated separately, each represented as a classification problem with different classes binning an angle range: We use 30, 13, and 30 classes for pitch, yaw, and roll, respectively. For objects with axial symmetry, the roll component is not estimated. Similarly, the 3D depth values are estimated via classification (into 140 bins). The standard soft-max cross-entropy loss is used for training (L_P for 3D pose and L_D for depth):

$$L_{P(D)} = \frac{1}{N} \sum_{i=1}^N \mathbf{y}_i^* \cdot \log(\mathbf{y}_i), \quad (3)$$

where \mathbf{y}^* is the ground-truth in the form of one-hot encoding, \mathbf{y}_i is the class estimation after softmax activation, and N is the batch size. Combining the hypotheses across all rotation and depth dimensions, we obtain 1.638M possible combinations (1.75K for axially symmetric objects) each provided with a confidence value as a product of individual classification confidences. These confidences are used by the joint registration module along with the spatial

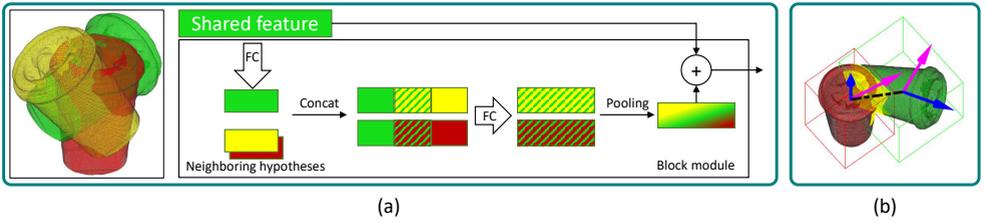


Figure 3: **(a)** Schematic diagram of the feature extractor in the joint registration module (inspired by [13]) applied to a hypothesis (green) with multiple neighboring hypotheses. Hatched boxes are the 6D relation features. Color in the box corresponds to hypotheses on the left image. Features from different hypothesis and 6D relation feature are concatenated. Variable number of neighbors for each hypothesis is reduced to a single feature by max-pooling. **(b)** An example neighborhood pattern: The intersection of the corresponding bounding boxes is highlighted in yellow; The black dotted line joints the center of bounding boxes; Blue and magenta arrows represent yaws and pitches, respectively.

configurations of neighboring detections. For computational efficiency, before joint registration, we sample only five hypotheses per object by first removing many similar hypotheses using NMS, and then selecting the remaining hypotheses with the five highest confidences.

Joint registration. By design, the pool of hypotheses generated from 6D pose estimation module contains multiple false positives. As no prior information on the number of ground-truth object instances or their spatial configurations is available, the joint registration module needs to consider simultaneously the spatial configurations of multiple hypothesized object poses. We adapt GossipNet framework proposed by Hosang *et al.* [13] for this purpose: Applied to the bin-picking scenario, our GossipNet adaptation classifies each hypothesis of interest (HOI) into True positive (1) or False positive (-1) classes. The classification is made based on a *contextual feature* encapsulating the spatial configurations of neighboring hypotheses (whose 3D bounding boxes overlap with the HOI). Figure 3(a) visualizes this feature extraction process: First, for a given HOI (green), all neighboring hypotheses (NH; e.g. red) are identified. Then, for each identified NH and HOI pair, an intermediate feature vector is generated by concatenating 1) the image depth features calculated within the bounding boxes of HOI and NH, respectively (solid green and red boxes in Fig. 3(a); See Fig. 1(shared feature)) and 2) 13-dimensional feature vector of their joint configurations(6D relation feature): (1-2) confidence score of each detection, (3) intersection over union score of bounding boxes, (4-9) Euler angles normalized to $[0, 1]$, (10-12) distances between the two detection centers along each x , y , and z direction, normalized by the bounding box diameter, and (13) normalized L_2 distance between the two detection centers.

The multiple intermediate pair-wise features (for all identified NHs) constructed in this way are then pooled to make a single contextual feature vector facilitating the classification independently of the number of neighboring hypotheses. During training, the registration network f_J (combining the contextual feature extractor and a fully connected network classifier) minimizes the classification loss given below [13]:

$$L_J(f_J) = \sum_{i=1}^N a_i \cdot \log(1 + \exp(-f(h_i) \cdot y_i)). \quad (4)$$

Here, the batch-size N is the number of total hypothesis $\{h_i\}$ in a single depth image, $y_i \in \{-1, 1\}$ represents the ground-truths, and a_i is a weight compensating the class imbalance decided at 1 and 16 for 1 and -1 classes, respectively.

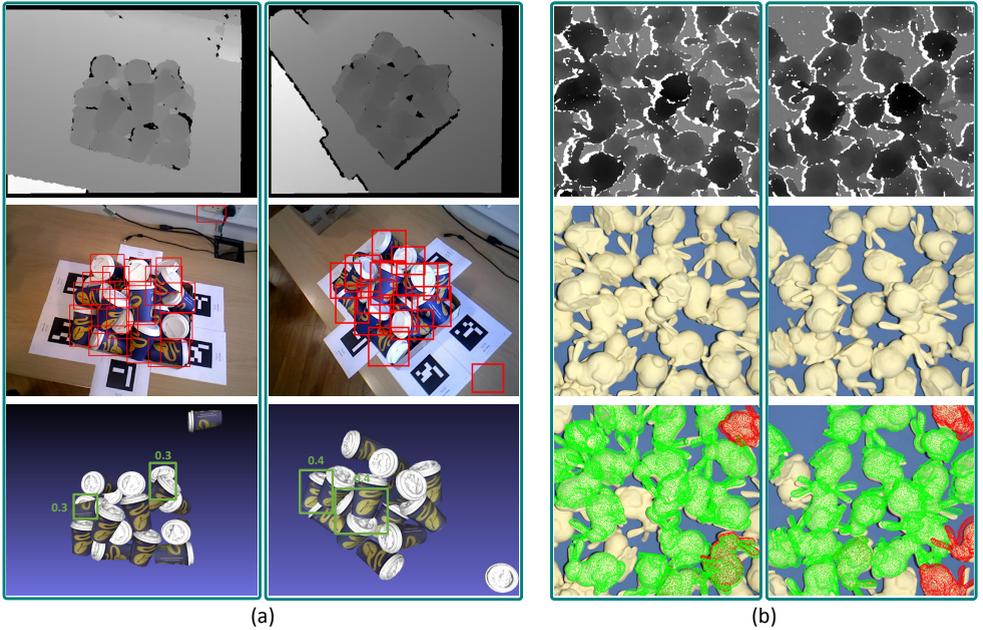


Figure 4: Qualitative results of top 20 hypotheses for the *Coffee cup* and *bunny* scenario. **(a)** For each triplet: (top) Input depth image. (middle) Detected bounding boxes. RGB image is used for visualization. (bottom) Reconstructed scene from hypotheses. Difficult objects are highlighted in green box annotated with visibility ratio. Objects with severe occlusion, some of which have visibility as low as 30%, are successfully recovered. **(b)** For each triplet: (top) and (middle) The input depth image and the corresponding RGB image (for visualization); (bottom) Hypotheses projected onto the RGB image. True and false positives are highlighted in green and red, respectively

4 Experiments

We evaluate our network on four objects from two different bin-picking datasets where multiple object instances occlude each other, reflecting realistic application cases: *Coffee cup* from the Bin-Picking dataset [10], and *Bunny*, *Pepper*, and *Candlestick* from the Siléane dataset [11]. We compare with existing algorithms: Linemod algorithm [12]; 3D point-pair-feature (PPF)-based algorithm [8]; and Latent-class Hough forest (LCHF) [13]. While PPF and LCHF are known to be robust to scene clutters and occlusions, respectively, they are not specifically designed for bin-picking applications. Therefore, we also compare with hypotheses verification approach [14] that extends Linemod and PPF for bin-picking applications (referred to as ‘Linemod+PP’ and ‘PPF+PP’, respectively). To gain an insight into the effectiveness of our multi-task learning approach in general 6D pose estimation (non bin-picking cases), we also evaluate our algorithm on the Domestic Environment dataset (*Coffee cup* object) [14] providing examples of multiple object instances under mild occlusions.

The accuracy of predictions is measured in average precision (AP; equivalently, the area under the precision-recall curve) evaluated based on *Sym* criteria adopted from [8]: Based on *Sym*, a hypothesis is accepted (considered as a true positive) if the distance to ground-truth is smaller than 10% of the diameter of the smallest bounding sphere. We calculated the AP values of our algorithm using the evaluation code shared by the authors of [8] enabling a

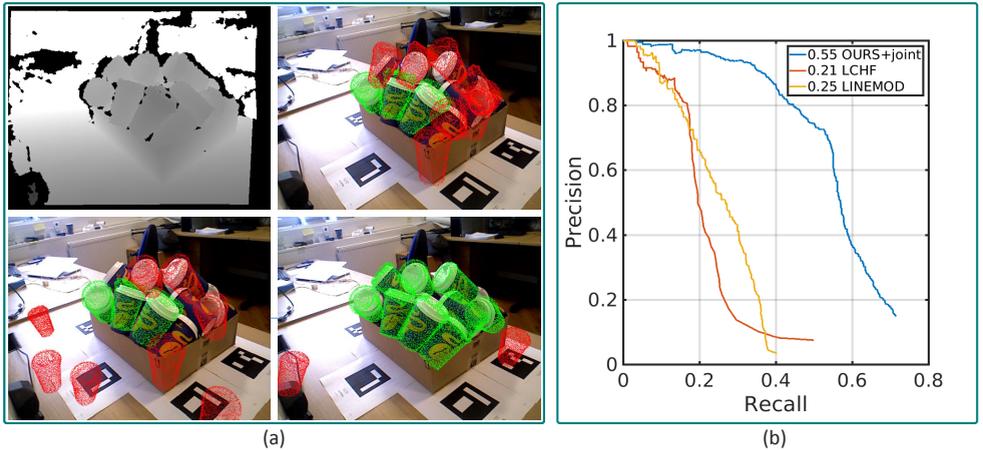


Figure 5: (a) Input depth image and example 6D pose estimates by Linemod, LCHF, and Ours, respectively (clockwise from top-left). Correctly estimated poses are highlighted in green. (b) The precision and recall curves of Linemod, LCHF, and our algorithm on *Coffee cup* dataset.

Table 1: (a-b) AP values of different algorithms in crowd scenario: Each pose hypothesis is evaluated in *Sym* criterion (a) and *ADD* criterion (b). (c) F1 scores for Domestic Environment dataset (*Cup* object) [14] with *ADD*.

Method	Bunny [14]	cup [14]	pepper [14]	C.stick [14]
Linemod [14]	0.39	0.08	0.04	0.38
Linemod+PP [14]	0.45	0.20	0.03	0.49
PPF [14]	0.29	0.28	0.06	0.16
PPF+PP [14]	0.37	0.30	0.12	0.22
Ours (w/o offset)	0.49	0.53	0.32	0.53
Ours	0.74	0.55	0.43	0.64

(a)

Method	Cup [14] (real)	Cup [14] (synt.)
Linemod [14]	0.38	0.32
LCHF [14]	0.30	0.31
SSD6D [14]	0.07	0.01
Ours	0.68	0.80

(b)

Method	Cup [14]
Linemod [14]	0.819
PPF [14]	0.867
LCHF [14]	0.877
Sahin et al. [14]	0.793
Kehl et al. [14]	0.972
Ours	0.992

(c)

direct comparison with the results of Linemod(+PP) and PPF(+PP) supplied from [14]. We also measure the AP based on *ADD* criteria [14] which accepts a hypothesis if its average distance from the ground-truth pose is less than 0.1 times the object diameter.

Results on bin-picking datasets.² Table 1(a) summarizes the AP scores. No clear winner was found among the two classical algorithms, Linemod and PPF: Each one is better than the other on two datasets. As these algorithms are not designed for heavily cluttered-scenes, applying extra verification [14] significantly improved performance (Linemod+PP and PPF+PP). None of these four existing methods accounts for the occlusion patterns during training. Instead, they use the ICP to verify and refine the estimated hypothesis. By explicitly and simultaneously learning the occlusion patterns and object shapes during training, our algorithm (Ours) constantly outperformed existing algorithm with a large margin (15-31%) even without having to rely on the external ICP algorithm.

For qualitative comparison, Fig. 5(a) shows example detections: LCHF, Linemod, and Ours all correctly estimated the 6D poses of unoccluded foreground objects. However, LCHF and Linemod missed multiple objects with only mild levels of occlusions (highlighted in

²As the original Bin-Picking dataset contained erroneous ground-truth annotation, every frame has been manually re-annotated via visual inspection.

red), while our algorithm correctly recovered all objects with only two false positives leading to constantly and significantly higher precision across different recall values (Fig. 5(b)).

Due to severe occlusions caused by multiple instances of a single object, estimating the 6D poses in bin-picking using pure RGB-based approaches is challenging. We verified this by applying the state-of-the-art SSD6D algorithm [16] in this scenario³ (Table 1): Not surprisingly, the results of SSD6D algorithm failed to match the level of any depth image-based pose estimation algorithms in comparison.

‘Ours (w/o offset)’ in Table 1(a) shows the AP score of our algorithm obtained without the object center estimation to assess its contribution: Under severe occlusion, the initial bounding box detections are not tightly aligned with the ground-truths. Therefore, excluding this component severely degraded the joint registration process.

We also evaluate the performance of our algorithm with ADD (Table 1(b)), which confirms the benefit of our explicit occlusion pattern learning approach. Note that both Linemod and LCHF, in comparison, use RGBD data whereas ours uses only depth images.

Non bin-picking scenario. We use the F1 score following publicized results supplied in [15, 28] (Table 1(c)): Even though our algorithm was trained for bin-packing scenes (with training data consisting of only foreground scenes), and it does not use the ICP refinement in contrast to others, it significantly outperformed Linemod [10], PPF [8], and LCHF [24] and is on par with the state-of-the-art kehl *et al.*’s algorithm [15] demonstrating the effectiveness and robustness (to unseen background clutters) of our algorithm.

Implementation details and complexities. Our shared feature extractor (Fig. 1) is initialized with Imagenet-trained VGG network [23]. The entire network was trained using the ADAM optimizer with 10^{-5} learning rate. Training the network on each object took around 10 hours in a machine with a GTX1080ti GPU. Testing a single depth image took 204ms.

5 Conclusion

Object detection and pose estimation in bin-picking scenarios is challenging due to severe occlusions caused by multiple instances of the same object. Existing approaches are limited in that their estimators are trained on isolated object instances and they do not directly consider possible mutual occlusion patterns during training. We presented a new 6D pose estimation framework that is trained directly on the scenes demonstrating multiple occlusion patterns. Our framework is instantiated as 2D detection, 6D pose hypothesis generation, and joint registration submodules which are jointly trained via a new multi-task loss. The resulting algorithm significantly outperformed the state-of-the-art 6D object pose estimation approaches on challenging datasets.

Acknowledgement: This work is part of Imperial College London-Samsung Research project, supported by Samsung Electronics. Kwang In Kim thanks RCUK EP/M023281/1.

³We used the code kindly shared by the authors of this work.

References

- [1] A. Aldoma, F. Tombari, L. D. Stefano, and M. Vincze. A global hypotheses verification method for 3D object recognition. In *ECCV*, 2012.
- [2] K. Alex and C. Roberto. Geometric loss functions for camera pose regression with deep learning. In *CVPR*, 2017.
- [3] V. Balntas, A. Doumanoglou, C. Sahin, J. Sock, R. Kouskouridas, and T.-K. Kim. Pose guided rgb-d feature learning for 3D object pose estimation. In *ICCV*, Oct 2017.
- [4] E. Brachmann, A. Krull, F. Michel, S. Gumhold, J. Shotton, and C. Rother. Learning 6d object pose estimation using 3D object coordinates. In *ECCV*, 2014.
- [5] R. Brégier, F. Devernay, L. Leyrit, and J.L. Crowley. Symmetry aware evaluation of 3D object detection and pose estimation in scenes of many parts in bulk. In *ICCV*, 2017.
- [6] T. T. Do, M. Cai, T. Pham, and I. D. Reid. Deep-6dpose: Recovering 6D object pose from a single RGB image. *arXiv:1802.10367*, 2018.
- [7] A. Doumanoglou, R. Kouskouridas, S. Malassiotis, and T.-K. Kim. Recovering 6D object pose and predicting next-best-view in the crowd. In *CVPR*, 2016.
- [8] B. Drost, M. Ulrich, N. Navab, and S. Ilic. Model globally, match locally: Efficient and robust 3D object recognition. In *CVPR*, 2010.
- [9] C. Eppner, S. Höfer, R. Jonschkowski, R. Martin, A. Sieverling, V. Wall, and O. Brock. Lessons from the amazon picking challenge: Four aspects of building robotic systems. In *Robotics: Science and Systems*, 2016.
- [10] S. Hinterstoisser, S. Holzer, C. Cagniart, S. Ilic, K. Konolige, N. Navab, and V. Lepetit. Multimodal templates for real-time detection of texture-less objects in heavily cluttered scenes. In *ICCV*, 2011.
- [11] S. Hinterstoisser, V. Lepetit, S. Ilic, S. Holzer, G. Bradski, K. Konolige, and N. Navab. Model based training, detection and pose estimation of texture-less 3D objects in heavily cluttered scenes. In *ACCV*, 2012.
- [12] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 1997.
- [13] J. Hosang, R. Benenson, and B. Schiele. Learning non-maximum suppression. In *CVPR*, 2017.
- [14] H. Hu, J. Gu, Z. Zhang, J. Dai, and Y. Wei. Relation networks for object detection. *arXiv:1711.11575*, 2017.
- [15] W. Kehl, F. Milletari, F. Tombari, S. Ilic, and N. Navab. Deep learning of local rgb-d patches for 3D object detection and 6D pose estimation. In *ECCV*, 2016.
- [16] W. Kehl, F. Manhardt, F. Tombari, S. Ilic, and N. Navab. SSD-6D: Making rgb-based 3D detection and 6D pose estimation great again. In *ICCV*, 2017.

- [17] A. Krull, E. Brachmann, F. Michel, M. Ying Yang, S. Gumhold, and C. Rother. Learning analysis-by-synthesis for 6d pose estimation in rgb-d images. In *ICCV*, 2015.
- [18] D. Lee, G. Cha, M-H Yang, and S. Oh. Individualness and determinantal point processes for pedestrian detection. In *ECCV*, 2016.
- [19] M-Y. Liu, O. Tuzel, A. Veeraraghavan, Y. Taguchi, TK Marks, and R. Chellappa. Fast object localization and pose estimation in heavy clutter for robotic bin picking. *IJRR*, 2012.
- [20] K. Park, J. Prankl, and M. Vincze. Mutual hypothesis verification for 6D pose estimation of natural objects. In *ICCV Workshops on Recovering 6D Object Pose*, 2017.
- [21] M. Rad and V. Lepetit. BB8: A scalable, accurate, robust to partial occlusion method for predicting the 3D poses of challenging objects without using depth. In *ICCV*, 2017.
- [22] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*, 2015.
- [23] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, and M. Bernstein. Imagenet large scale visual recognition challenge. *IJCV*, 2015.
- [24] C. Sahin, R. Kouskouridas, and T.-K. Kim. A learning-based variable size part extraction architecture for 6D object pose recovery in depth. In *IVC*, 2017.
- [25] J. Sock, S. Hamidreza Kasaei, L. Seabra Lopes, and T.-K. Kim. Multi-view 6D object pose estimation and camera motion planning using rgb-d images. In *ICCV Workshops on Recovering 6D Object Pose*, 2017.
- [26] R. Stewart, M. Andriluka, and A.Y. Ng. End-to-end people detection in crowded scenes. In *CVPR*, 2016.
- [27] A. Tejani, D. Tang, R. Kouskouridas, and T.-K. Kim. Latent-class hough forests for 3D object detection and pose estimation. In *ECCV*, 2014.
- [28] A. Tejani, R. Kouskouridas, A. Doumanoglou, D. Tang, and T. K. Kim. Latent-class hough forests for 6DoF object pose estimation. *TPAMI*, 2018.
- [29] B. Tekin, S. Sinha, and P. Fua. Real-time seamless single shot 6D object pose prediction. *arXiv preprint arXiv:1711.08848*, 2017.
- [30] K. Walas, T. Hodan, T.K. Kim, J. Matas, C. Rother, F. Michel, V. Lepetit, A. Leonardis, C. Stegger, C. Sahin, and R. Kouskouridas. 3rd international workshop on recovering 6D object pose, 2017. URL http://cmp.felk.cvut.cz/sixd/workshop_2017/.
- [31] Y. Xiang, T. Schmidt, V. Narayanan, and D. Fox. Posecnn: A convolutional neural network for 6D object pose estimation in cluttered scenes. *arXiv:1711.00199*, 2017.
- [32] A. Zeng, K-T. Yu, S. Song, D. Suo, E. Walker, A. Rodriguez, and J. Xiao. Multi-view self-supervised deep learning for 6D pose estimation in the amazon picking challenge. In *ICRA*, 2017.