Pattern Recognition 154 (2024) 110576

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

# OA-Pose: Occlusion-aware monocular 6-DoF object pose estimation under geometry alignment for robot manipulation

Jikun Wang [a], Luqing Luo [b], Weixiang Liang [a], Zhi-Xin Yang [a,*]

[a] The State Key Laboratory of Internet of Things for Smart City, Department of Electromechanical Engineering and Centre of Artificial Intelligence and Robotics, University of Macau, 999078, Macao Special Administrative Region of China

[b] The Institute of Microelectronics, Chinese Academy of Sciences, Beijing, 100029, China

ARTICLE INFO

*Keywords:*
Object pose estimation
Occlusion scene
Deep learning
Dense correspondence
Robot manipulation

ABSTRACT

Object pose estimation is the fundamental technology of robot manipulation systems. Recently, various learning-based monocular pose estimation methods have achieved outstanding performance by establishing sparse/dense 2D–3D correspondences. However, in cluttered environments, occlusion has been a challenging problem for pose estimation due to limited information provided by visible parts. In this work, we propose an efficient occlusion-aware monocular pose estimation method, called OA-Pose, to learn geometric feature information of occluded objects from cluttered scenes. Our framework takes RGB images as input and generates 2D–3D correspondences of visible and invisible parts based on the proposed Occlusion-aware Geometry Alignment Module. Extensive experiments show that our method is superior and competitive with state-of-the-art on multiple public datasets. We also conduct grasping experiments with different degrees of object occlusion, demonstrating the usability of our algorithm to deploy on robots in unstructured environments.

## 1. Introduction

With the development of smart city intelligent technology, industrial robots have been applied in many fields [1]. Object pose estimation plays an important role in obtaining pose parameters (rotation **R** and translation **t**) to prepare for downstream robotic manipulation (Fig. 1) [2]. Many learning-based pose estimation algorithms perform robustly from monocular color images in neat and clean scenes. However, in real-world applications, robots are demanded to work in unstructured environments with objects placed in arbitrary illumination conditions under cluttered scenes [3]. The occluded objects in images provide extremely limited information, making accurate pose estimation challenging.

Some traditional methods accomplish pose estimation by combining with depth data [4]. However, effective depth data is not easy to obtain [5]. Some vision sensors and augmented reality do not collect depth data. With the prevalence of deep learning technologies, both the performance and accuracy of monocular pose estimation have been advanced. Therefore, plenty of research focuses on pose estimation from RGB images. Recently, monocular pose estimation method is roughly classified into three categories that include the one-stage method, the two-stage method and the differentiable two-stage method. The one-stage method is to directly regress the 3D rotation and 3D translation in an end-to-end fashion. Since direct regression of pose parameters is a difficult task, there are many research directions for this problem in one-stage methods. Posecnn [6] and DeepIM [7] design point matching loss to enhance model training. Manhardt et al. [8] divides the pose space into different discrete spaces, that means converting the continuous space matrix regression problem into a classification problem, which reduces the difficulty of the one-stage method. PoET [9] can directly predict the pose parameters of the object by using the designed transformer decoder. Nevertheless, the one-stage methods normally are inferior to the two-stage ones on challenging scenarios, such as occlusion, symmetric objects, etc. The two-stage method is to utilize a two-stage strategy, where 2D–3D correspondences are estimated first, and pose parameters are acquired through post-processing such as RANSAC-based Perspective-n-Point (PnP) algorithms. 2D–3D correspondence is divided into key point prediction, corner point prediction and dense prediction. PVNet [10] uses predefined key points from object surface as 2D–3D correspondence to predict pose parameters. Setting key points manually is time-consuming. Tekin et al. [11] predicts the corner points of the object's 3D bounding box as 2D–3D correspondence. Bdr6d [12] predicts key points by the designed depth estimation network and the proposed bidirectional deep residual (BDR) fusion network, and finally calculates pose parameters using the PnP algorithm. However, both keypoint prediction and corner point prediction are sparse, and model performance is insufficient. Therefore,

* Corresponding author.
*E-mail addresses:* yc07492@um.edu.mo (J. Wang), luoluqing@ime.ac.cn (L. Luo), yc27951@um.edu.mo (W. Liang), zxyang@um.edu.mo (Z. Yang).

https://doi.org/10.1016/j.patcog.2024.110576
Received 20 February 2023; Received in revised form 9 April 2024; Accepted 8 May 2024
Available online 14 May 2024

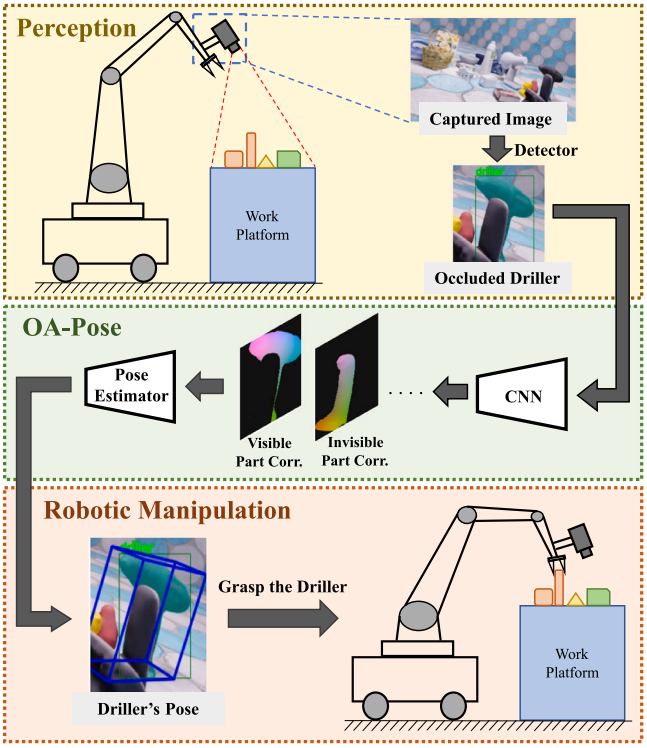


**Fig. 1.** Illustration of robotic manipulation with OA-Pose. After an image is captured by the vision sensor (Yellow) using the two-branch network of OA-Pose to predict the 2D–3D correspondence of visible and invisible parts, pose parameters can be regressed directly (Green). The robot can thus perform corresponding manufacturing operations based on the estimated poses of the workpieces (Orange). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

CDPN [13] and ER-Pose [14] improve model accuracy by predicting the dense 2D–3D correspondences. Aing et al. [15] also proposed a novel method to generate dense 2D–3D correspondence, and finally using the PnP algorithm to calculate the pose parameters. The two-stage strategy usually performs well with the post-processing optimization. But the extra post-refinement (RANSAC iterative process for computing pose parameters) is time-consuming when processing dense 2D–3D correspondence and degrades the efficiency of robotic applications.

The differentiable two-stage method combines the high accuracy of the two-stage strategy and the end-to-end training mode of the one-stage strategy. This strategy improves the algorithm efficiency of the original two-stage method by designing a learnable PnPNet. GDR-Net [16] couples 2D–3D correspondence into the proposed network and predicts pose parameters based on the end-to-end structure. RNNPose [17] designs the recursive neural network to optimize poses during the iterative process. For the differentiable two-stage method, the predicted dense 2D–3D correspondences contain more geometric information, so this method has better performance, which also makes it difficult for the model to predict intermediate geometric features [18]. Furthermore, when there are occluded objects in the image, the absence of invisible parts makes it more difficult to establish an effective dense 2D–3D correspondence [19]. To address this problem, we build additional supervision for the invisible parts based on the geometric information alignment, thereby improving the accuracy of predicted correspondences for occluded objects. In this work, we propose an end-to-end Occlusion-Aware monocular 6-DoF pose estimation network (OA-Pose) based on the differentiable two-stage strategy to regress 6D pose parameters according to occlusion-aware geometric features. Concretely, the occlusion-aware geometric features are determined through 2D alignment between object coordinates and target region of image pixels and 3D alignment between camera and object coordinate systems. The coordinate-based geometry alignment benefits the pose estimation by driving the invisible correspondences toward the occluded parts of objects consistent across 2D and 3D spaces, as shown in Fig. 1. Compared to state-of-the-arts, our method shows the competitive performance and the grasping experiments on robots demonstrate its veracity and generality. The key contributions are summarized below:

- New occlusion-aware geometry alignment method is proposed to predict accurate and dense pixel-level 2D–3D correspondence, and thus improve the pose estimation performance in occlusion scenes.
- The proposed loss functions $L_{mi-2D}$ and $L_{mc-3D}$ drive the invisible correspondences toward the occluded parts of objects consistent across the 2D and 3D spaces.
- The proposed OA-Pose not only achieves state-of-the-art performance on various challenging public datasets, but also demonstrates successful industrial applications in real and complex robot grasping tasks.

The following sections are organized as following, the related works are briefly reviewed in Section 2, the details of systematic framework of OA-Pose are explained in Section 3. In Section 4, the proposed method is evaluated on different datasets. Finally, Section 5 concludes the paper.

## 2. Related works

Object 6D pose estimation is a challenging task that is to estimate the 3D pose in 2-tuple geometric transformations (rotation and translation) of the object with input data in RGB, depth data or RGB-D formats. Pose estimation methods based on depth or RGB-D data, such as [20,21], can achieve accurate and robust object tracking. However, when depth data cannot be obtained, monocular pose estimation methods are effective and feasible. Monocular pose estimation algorithm takes a single RGB data as input and estimates the 6D pose parameters. In this section, based on the algorithm structure, we review the monocular pose estimation algorithms with different strategies. Some methods utilize the one-stage method to estimate the final pose directly from a single image. The other is to build the two-stage method that first predicts intermediate geometric features and then estimates pose parameters. Furthermore, the third is to design the differentiable two-stage method to finish pose estimation based on end-to-end training mode and 2D–3D correspondence.

Compared with the two-stage method, the end-to-end network structure is easier to apply to a variety of tasks. Therefore, some methods [6,7] use point matching loss to directly regress the pose parameters. For some methods [8], the pose space is divided into discrete spaces, and the pose is obtained by model classification instead of regression algorithm [22]. In this case, pose regression is transformed into a classification problem. In [23], using pose refinement to improve end-to-end network performance. Based on the proposed bidirectional feature pyramid network (BiFPN), the EfficientPose [24] introduces an efficient method for 6D pose estimation. PoET [9] design a decoder with transformer as the core to predict the pose parameters of the object only using RGB image. Furthermore, the robustness of these methods is improved by adding multiple hypotheses. Direct regression methods are end-to-end models, however, as demonstrated in [25], this strategy has limitations in accuracy.

Another approach is to estimate pose parameters by establishing a two-stage strategy. In [10], the 2D–3D correspondence is established by predicting pre-set keypoints on the surface of object. Hybrid-Pose [26] further extends [10] by introducing hybrid representations. However, keypoint-based methods need to set several keypoints of objects. Furthermore, these keypoints are sparse, which is insufficient to perform complete object information. In this case, in [11], the 2D–3D correspondence is established by predicting the corners of the object's

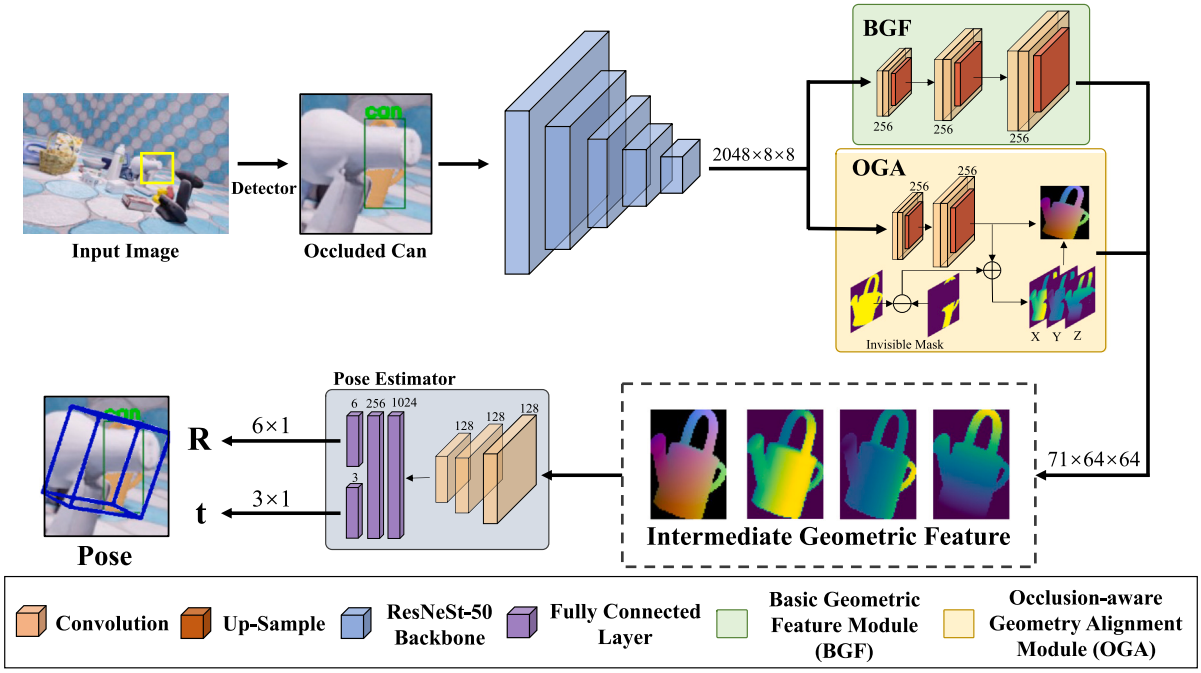


**Fig. 2.** Framework of the proposed OA-Pose. With the image as input, firstly, we use a detector to crop out the target object, and high-level features are extracted via the network backbone to feed in the subsequent decoders. The first decoder obtains the basic geometric features of the object (BGF), and the second one acquires the pixel-level 2D–3D correspondence based on Occlusion-aware Geometry Alignment Module (OGA). Finally, the intermediate geometric features obtained by two decoders are sent to the pose estimation part, and the 6D parameters are directly regressed.

3D bounding box, which greatly reduces the training cost. However, since eight corner points are located outside the model, so the wrong object position may be obtained, which greatly limits the accuracy of pose prediction. To enhance the stability of the network, in [10], a voting mechanism is added to each correspondence, but it is still sparse. MLFNet [27] builds a predefined code-book, then predicts the potential embedding of the pose through the network, and finally retrieves the pose by comparing the code-book and the embedding. Some works [13,14,28] predict the representation of the three-dimensional coordinates corresponding to each pixel, and then use the RANSAC-based PnP algorithm to estimate the pose parameters. Aing et al. [15] proposed a novel diagonal graph clustering to generate 2D–3D correspondence, and finally combined with the PnP RANSAC algorithm to calculate the pose parameters. Bdr6d [12] predicts the key points of the object based on the proposed depth estimation network and bidirectional deep residual (BDR) fusion network, and then calculates pose parameters using the PnP algorithm. Post-processing can improve the accuracy, however, the 2D–3D correspondence of the two-stage strategy is dense, and a large number of parameters leads to a decrease in the efficiency of the algorithm [29].

Recent studies have proposed the differentiable two-stage method by combining the advantages of high accuracy of the two-stage method and the characteristics of the end-to-end training of the one-stage method. These methods first predict the 2D–3D correspondence, and then use the designed learnable PnP network to predict the pose parameters [17,30]. The predicted intermediate geometric features (2D–3D correspondence) are dense correspondence, which needs to know the model geometry in advance [16,18]. Different from sparse correspondence, when computing dense correspondence, it is necessary to consider the geometric relationship of all model points. However, although there have been studies on the pose estimation of occluded objects [19], the estimation of the geometric information in the occluded parts is still challenging. Therefore, it is challenging and important to predict comprehensive and accurate correspondence from limited RGB information.

## 3. Methodology

### 3.1. Problem definition

The learning-based monocular pose estimation algorithm utilizes RGB images as input to map the pose parameters **R** and **t** from the object coordinate system to the camera coordinate system via a learned model $\mathcal{M}$ with trainable parameter $\mathcal{W}$.

$$\mathbf{R}, \mathbf{t} = \mathcal{M}\left(\mathcal{I}_{RGB}; \mathcal{W}\right), \tag{1}$$

The monocular pose estimation method based on dense correspondence mainly depends on the visible part of the object in the image. However, when there exhibits severe occlusions, the accuracy of the predicted geometric features deteriorates. Due to the limited information provided in cluttered environments, it is far from trivial to regress the accurate pose in light of the predicted geometric features.

Different from the previous dense correspondence-based pose estimation methods, in this paper, we present a novel occlusion-aware pose estimation method, OA-Pose, to handle challenging robot manipulation scenes of occlusions by using a single RGB image. In Fig. 2, given an image with occluded objects as input, the proposed OA-Pose puts forward a coordinate-based geometry alignment for invisible correspondences, and the acquired discriminate occlusion-aware geometric features contribute to achieve effective pose estimation.

As described in Fig. 3, points E-e and D-d are from the visible parts, and points A-a, B-b and C-c are from the invisible parts. The features drawn from the visible part of occluded objects normally struggle in perceiving overall geometric information since the invisible part could contain important semantic information which is difficult to retrieve.

### 3.2. Occlusion-aware geometry alignment module

We devise an Occlusion-aware Geometry Alignment Module (OGA) to establish the geometric feature-based correspondence of occluded objects and reconstruct the enriched object geometric information for a direct pose regression. As shown in Fig. 4, by improving the accuracy of

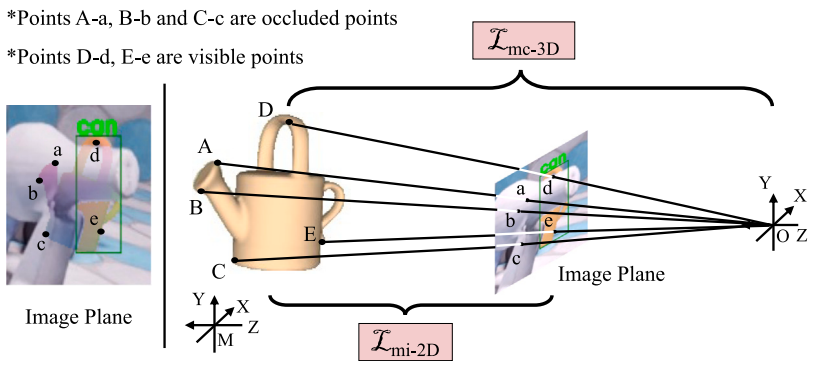


**Fig. 3.** Perspective projection transformation. Given the occluded object image, the limited visible part is used to predict the 2D–3D correspondence of the object, and finally the pose parameters are obtained. Here, we show the projection transformation relationship between the image plane, the camera coordinate system, and the object coordinate system. The object coordinate system and the camera coordinate system are denoted by M and O, respectively. Points A-a, B-b and C-c are the occluded parts. Points D-d and E-e are visible parts.

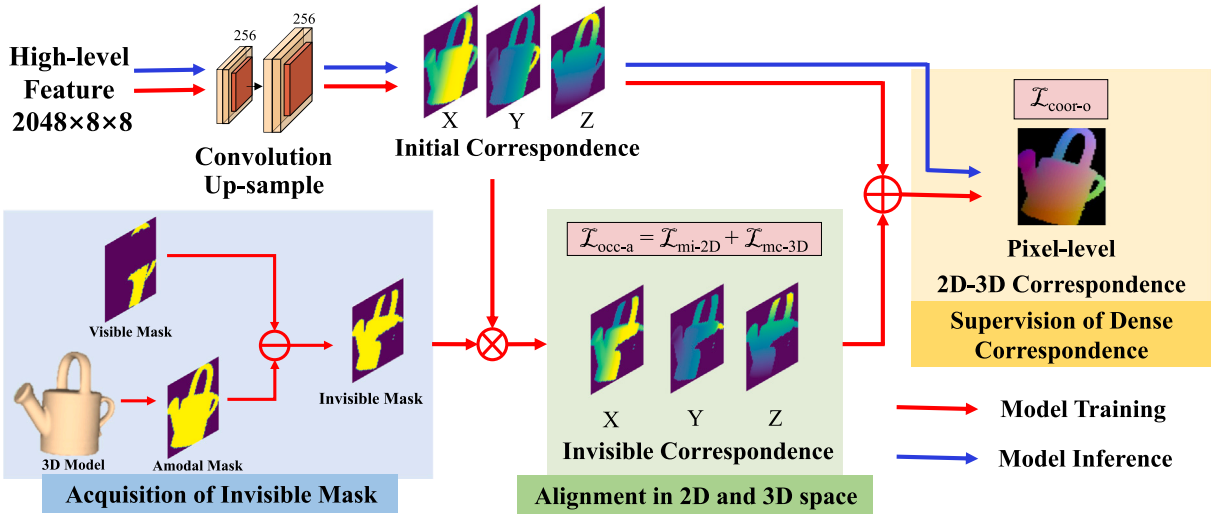


**Fig. 4.** Illustration of the Occlusion-aware Geometry Alignment Module. By improving the accuracy of occluded correspondence, pixel-level 2D–3D correspondence is obtained.

occluded correspondence, the predicted pixel-level 2D–3D correspondence is obtained. First, the initial correspondence is predicted based on the high-level features from four convolution layers (each followed by Batch Normalization and ReLU activation) and two up-sample layers. Then, the invisible mask of the occluded part is calculated by the visible mask from the data set and the amodal mask from the 3D model. Then, based on the proposed $L_{mi-2D}$ and $L_{mc-3D}$, the invisible correspondence is strengthened. Finally, the initial correspondence and invisible correspondence are merged to obtain the occlusion-aware pixel-level 2D–3D correspondence. In the following, we detail the acquisition of occlusion-aware geometric features and the derivation of the occlusion-aware loss functions.

Firstly, we utilize the perspective projection transformation to acquire dense correspondence. The derived geometric feature-based representations from the invisible parts are used to align the occluded parts according to the geometrical consistency across 3D and 2D spaces. To establish the dense correspondence, 3D points of the object model are projected onto 2D plane, the transformation can be written as,

$$\mathcal{I}_i = P_{ci}\left(K, Z_{D_c}; P_{mc}\left(D_m, \hat{\mathbf{R}}, \hat{\mathbf{t}}\right)\right), \tag{2}$$

where $D_c$ is the 3D point in camera coordinate system and $\mathcal{I}_i$ is the 2D point in image plane. Combining 3D point coordinate $D_c$ with the image plane constitutes the 2D–3D correspondence as follow,

$$C_{gt} = \{D_m\}_i, \tag{3}$$

where $C_{gt} \in \mathbb{R}^3$ denotes the ground-truth correspondence, $i$ is the image plane, $K$ is camera intrinsic matrix, $Z_{D_c}$ is the Z value of the 3D point in the camera coordinate system, $P_{ci}(\cdot)$ and $P_{mc}(\cdot)$ are the projection of the camera coordinate system to the image plane (2D space) and the transformation of the object coordinate system to the camera coordinate system (3D space), respectively.

Secondly, based on the spatial relationship of generated correspondence (Eq. (2)), coordinate-based feature alignment loss functions in 3D and 2D spaces are constructed. Primarily, the transformation $P_{mc}(\cdot)$ of a 3D point from object coordinate system $D_m$ to camera coordinate system $D_c$ is:

$$D_c = \mathbf{R}D_m + \mathbf{t}, \tag{4}$$

where $\mathbf{R}$ and $\mathbf{t}$ represent 3D rotation and 3D translation, respectively. Specifically, $D_m = [X_{D_m}, Y_{D_m}, Z_{D_m}]^\top$, and $D_c = [X_{D_c}, Y_{D_c}, Z_{D_c}]^\top$. Moreover, the 3D points of camera coordinate system $D_c$ is projected onto the image plane via,

$$d = \frac{1}{Z_{D_c}} K D_c, \tag{5}$$

where $d = [x_d, y_d]$ is the 2D point of image plane.

Consequently, the projection $P_{ci}(\cdot)$ of object coordinate system to image plane (2D space) can be acquired by substituting Eq. (4) into

Eq. (5) as,

$$\frac{1}{Z_{D_c}} K \left( \mathbf{R} D_m + \mathbf{t} \right) - d = 0, \tag{6}$$

the illustration is shown in Fig. 3.

In 2D space, invisible points B-b, C-c and E-e respectively lie on the same ray. In order to maintain the consistency of 3D points in camera coordinate system and 2D points in image plane, a loss function $L_{mi-2D}$ is proposed to align the occluded parts between the object coordinate system and the image plane as,

$$\begin{aligned} L_{mi-2D} = & \sum_{D_{m_{\text{in}}} \in M} \left\| \frac{1}{Z_{D_c}} K \left( \mathbf{R} \hat{D}_{m_{\text{in}}} + \mathbf{t} \right) - \frac{1}{Z_{D_c}} K \left( \mathbf{R} D_{m_{\text{in}}} + \mathbf{t} \right) \right\|_1 \\ & + \sum_{D_{m_{\text{in}}} \in M} \left\| \frac{1}{Z_{D_c}} K \left( \mathbf{R} D_{m_{\text{in}}} + \mathbf{t} \right) - \hat{d} \right\|_1 \\ & + \sum_{D_{m_{\text{in}}} \in M} \left\| \frac{1}{Z_{D_c}} K \left( \hat{\mathbf{R}} D_{m_{\text{in}}} + \hat{\mathbf{t}} \right) - \hat{d} \right\|_1 , \end{aligned} \tag{7}$$

where $D_{m_{\text{in}}}$ is the 3D points of invisible parts in the object coordinate system whose range is extracted by $\hat{D}_m$ and invisible mask. M is the object model. $\hat{\mathbf{R}}$ and $\hat{\mathbf{t}}$ represent the ground truth. $\hat{d}$ is the ground truth, which represents the 2D point on image plane.

Furthermore, in 3D space, the 3D points on the object model belong to the same point in the camera coordinate system and the object coordinate system. In this case, consistency in 3D space is also imposed. According to Eq. (4), the transformation of the object coordinate system to the camera coordinate system is,

$$D_m = \mathbf{R}^\top D_c - \mathbf{R}^\top \mathbf{t}. \tag{8}$$

By multiplying $D_c$ on the equation above, we have,

$$D_m D_c + \mathbf{R}^\top \mathbf{t} D_c - \mathbf{R}^\top D_c D_c = 0. \tag{9}$$

By substituting Eq. (4) into Eq. (9), the alignment of occluded parts between the object coordinate system and the camera coordinate system can be imposed via,

$$\begin{aligned} L_{mc-3D} = & \sum_{D_{m_{\text{in}}} \in M} \| D_{m_{\text{in}}} \left( \mathbf{R} D_{m_{\text{in}}} + \mathbf{t} \right) + \mathbf{R}^\top \mathbf{t} \left( \mathbf{R} D_{m_{\text{in}}} + \mathbf{t} \right) \\ & - \mathbf{R}^\top \left( \mathbf{R} D_{m_{\text{in}}} + \mathbf{t} \right) \left( \mathbf{R} D_{m_{\text{in}}} + \mathbf{t} \right) \|_1 . \end{aligned} \tag{10}$$

Noted that the information of the occluded parts comes from the invisible mask, which is generated by the difference between the amodal mask and the visible mask. To this end, the alignment of invisible part in the correspondence field and pose parameter field is achieved. Furthermore, to improve the convergence efficiency of the proposed loss function, invisible mask is used during model training. During model inference, invisible masks do not participate in the prediction of pose parameters, as shown in Fig. 4.

### 3.3. Network architecture

The overall network architecture of OA-Pose is illustrated in Fig. 2, which is a two-branch framework predicting both basic geometric features and enhanced occluded correspondences. Here, the proposed Occlusion-aware Geometry Alignment Module (OGA) constructs the enhanced occluded correspondences branch, and the basic geometric feature module (BGF) tackles with visible part of objects by following the architecture of [16] with minor modifications. Overall, OA-Pose generates geometry information of the complete part to estimate 6D pose parameters.

The proposed method focuses on the 6D pose estimation of objects by taking advantage of the high accuracy and high efficiency of existing object detection algorithms [31] Therefore, same as [13,16,18,32], we decouple the training of the proposed model and the object detector. During training, we use zoomed-in dynamically to process the input image to ensure that the model can adapt to objects of different sizes.

Based on the input image I, we first detect objects of interest in the input data by using existing object detectors [1]. Then the cropped images are zoomed-in dynamically [13] and cropped to size 256 × 256 to extract high-level features via encoder ResNeSt-50 [33]. The intermediate geometric features of dimension 64 × 64 are generated by the designed decoder. The Occlusion-aware Geometry Alignment Module strengthen the consistency of three-channel occluded correspondence. In particular, based on the ground truth of pose parameters and 3D model, the occlusion correspondences are obtained with respect to perspective projection transformation from 3D model to the image plane. The Basic Geometric Feature Module (BGF) consists of six convolution layers (each followed by Batch Normalization and ReLU activation) and three up-sample layers. Similar to current differentiable two-stage pose estimation method, the Basic Geometry Module mainly predicts the feature of the visible part of the object and provides basic geometric information for the prediction of pose parameters. Based on the high-level features extracted from the RGB image, this module obtains geometric information including visible part 2D–3D correspondence, visible part mask, and object region feature. Here, amodal masks are employed for a complete information perception. The amodal mask is only used in the training phase, which is the complete mask of the object. Thereby, the proposed occlusion-aware geometry alignment module predicts the pixel-level 2D–3D correspondence of the complete object (Fig. 4). The Pose Estimator contains three convolution layers (each followed by Group Normalization and ReLU activation) for feature extraction and three fully connected layers to predict rotation parameters and translation parameters. Therefore, the geometric representations from both the visible and invisible parts are fed into the pose estimator [22] for 6D pose parameter regression, where $R_{6d}$ is used to parameterize the 3D rotation and Scale-Invariant representation for Translation Estimation (SITE) [13] is utilized for 3D translation.

### 3.4. Loss functions

The overall loss functions including basic pose terms and geometry alignment terms are expressed as,

$$L_{\text{all}} = L_p + L_{\text{occ-a}} + L_{\text{coor-o}}, \tag{11}$$

with

$$L_{\text{occ-a}} = L_{mi-2D} + L_{mc-3D}, \tag{12}$$

and

$$L_{\text{coor-o}} = \sum_{\mathcal{P}_m \in M} \left\| \mathcal{P}_m - C_{gt} \right\|_1 . \tag{13}$$

where $L_{\text{occ-a}}$ is the consistency term to perform additional geometric feature supervision on invisible parts of the object, $L_{\text{coor-o}}$ is the supervision of the 2D–3D correspondence in Occlusion-aware Geometry Alignment Module with $C_{gt}$ as the ground truth and $L_p$ is the loss function including correspondence, mask, region, 3D rotation and 3D translation as in [16]. $\mathcal{P}_m$ is the predicted pixel-level 2D–3D correspondence, which is the set of points $D_m$ in the object coordinate system.

In this way, the geometric features of complete object including visible and invisible parts are aligned in the correspondence field and pose parameter field. Although these geometric features are estimated as intermediate representations, the model is trained in an end-to-end fashion.

## 4. Experiment

This section presents experiments to evaluate our proposed pose estimation method. We conduct extensive experiments on two typical public datasets to demonstrate the superiority of our method, as well as grasp experiments to show the feasibility of robotics applications.

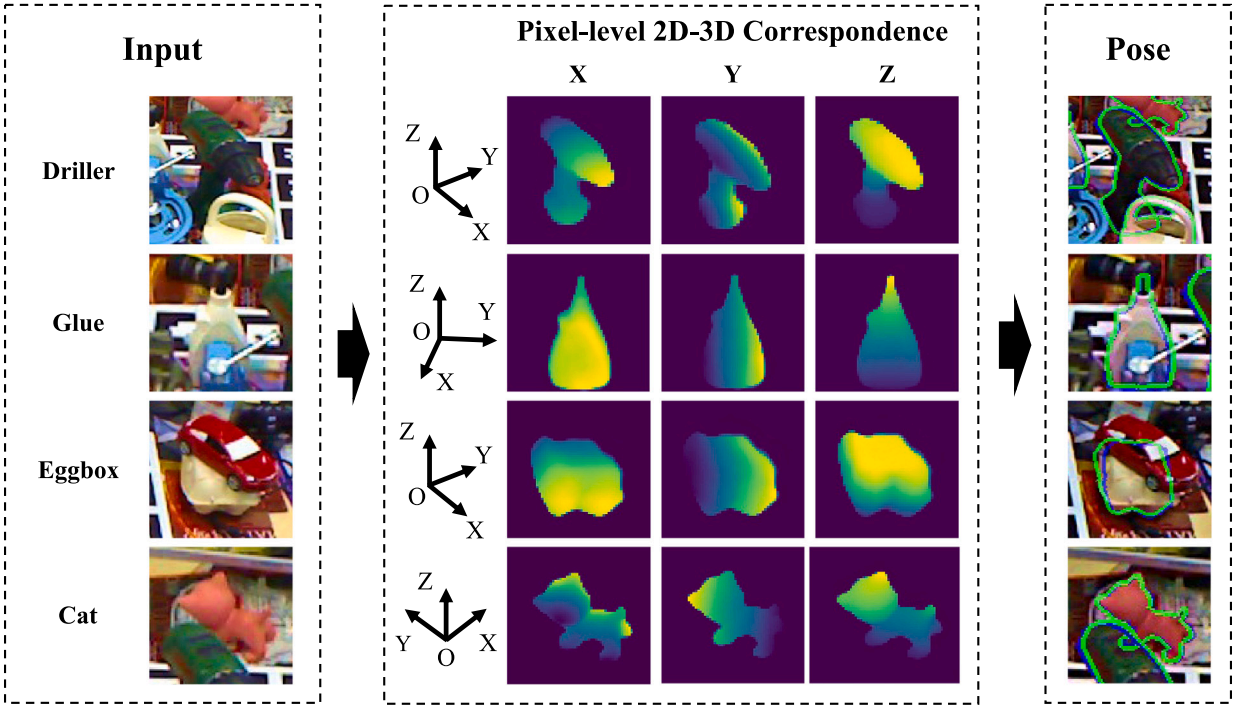


**Fig. 5.** Visualization results on LM-O. Based on the RGB image of the occluded object, the pixel level 2D–3D correspondence of the complete object is predicted, then the pose parameters are calculated. In *Pose* images, the *Blue* and *Green* Mask represent the ground truth 6D pose and the predicted results, respectively. It can be seen that the pose of the occluded objects(*Driller*, *Glue*, *Eggbox* and *Cat*) can be well estimated. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

**Table 1**
Results on LM referring to the Average Recall (%) of ADD(-S) metric.

| Object | w/o Refinement | | | | | | | w/ Refinement |
|---|---|---|---|---|---|---|---|---|
| | PoseCNN [6] | CDPN [13] | ER-Pose [14] | GDR-Net [16] | SO-Pose [18] | BDR6D [12] | Ours | RNNPose [17] |
| Ape | 25.62 | 67.33 | 62.6 | 76.29 | 85.43 | 90.6 | **96.59** | 88.19 |
| Benchvise | 77.11 | 98.74 | 100 | 97.96 | 99.42 | 98.2 | **100.0** | 100.0 |
| Cam | 47.25 | 92.84 | 95.8 | 95.29 | 96.67 | 98.5 | **99.21** | 98.04 |
| Can | 69.98 | 96.56 | 99.2 | 98.03 | 98.62 | 99.1 | **100.0** | 99.31 |
| Cat | 56.09 | 86.63 | 90.7 | 93.21 | 95.01 | 97.6 | **100.0** | 96.41 |
| Driller | 64.92 | 95.14 | 99.0 | 97.72 | 98.41 | 98.6 | 98.42 | **99.70** |
| Duck | 41.78 | 75.21 | 68.6 | 80.28 | 85.73 | 94.5 | **95.16** | 89.30 |
| Eggbox | 98.50 | 99.62 | 100 | 99.53 | 99.91 | 99.9 | 97.13 | **99.53** |
| Glue | 94.98 | 99.61 | 98.7 | 98.94 | 99.61 | 98.3 | **100.0** | 99.71 |
| Holepuncher | 52.24 | 89.72 | 89.7 | 91.15 | 94.77 | 92.7 | **100.0** | 97.43 |
| Iron | 70.17 | 97.85 | 99.6 | 98.06 | 98.67 | 98.4 | 96.44 | **100.0** |
| Lamp | 70.73 | 97.79 | 99.4 | 99.14 | 99.14 | 97.3 | 98.56 | **99.81** |
| Phone | 53.07 | 90.65 | 96.8 | 92.35 | 95.28 | 94.8 | 98.37 | **98.39** |
| Mean | 63.26 | 91.36 | 92.3 | 93.69 | 95.90 | 96.8 | **98.45** | 97.37 |

### 4.1. Datasets

We conduct experiments on three public datasets, LM (Linemod) [34], LM-O (Linemod-Occluded) [35] and YCB-V [6]. LM is the standard benchmark for 6D pose estimation of 13 objects in slightly cluttered scenes. Similarity to [16,18], 15% of the images are used for training and the rest are used for testing. Each image overlays the upper-view hemisphere on a cluttered table at a different scale. At the same time, each type of object is supplemented with additional 1k rendered images as the training set. LM-O is proposed based on LM, which contains 8 objects and is collected in severely occluded scenes. YCB-V is a large dataset including clutter and symmetric objects. This dataset has 21 categories of objects and a total of 110k real images. Meanwhile, for LM-O and YCB-V, publicly available synthetic datasets based on physically-based rendering [36] are used to supplement the training set.

### 4.2. Implementation details and image augmentation

Our network is trained on the Ranger optimizer with a learning rate of 1e–4 and a batch size of 36. In addition, to improve the stability of the proposed model, $L_{mc-3D}$ and $L_{mi-2D}$ are not added to the training initially and are added into the total loss function after 50% of the total training epoch. We initially use $L_{coor-o}$ for training. All experiments are implemented using PyTorch on RTX 2080Ti. To improve the generalization ability of the model, various random color enhancement strategies [16] are adopted. Furthermore, Faster RCNN [31] is used for 2D localization.

### 4.3. Evaluation metrics

We use the metrics ($n$ *cm*, $n°$ [37], ADD, ADD-S [34], Projection 2D [7] and BOP settings [38].) to evaluate performance. For $n$ *cm*, $n°$ [37], a pose with translation and rotation errors less than $n$ *cm*

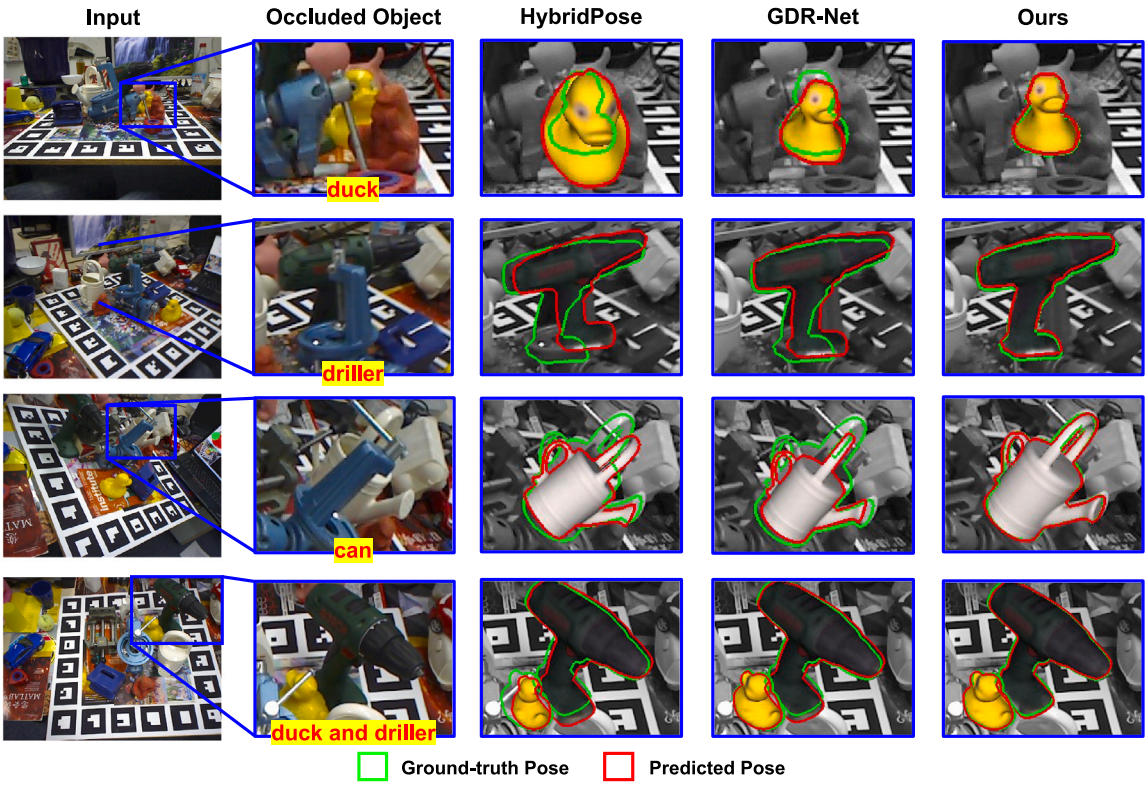


**Fig. 6.** Visual comparison of pose estimation in occlusion scenes on LM-O dataset.

and $n°$ is considered correct. For ordinary objects, ADD [34] means that the average error caused by the transformed model is less than 10% of the object diameter (0.1d). For symmetric objects, the ADD-S [34] is employed to calculate the distance to the closest model point. In Experiment section, ADD and ADD-S are collectively referred to as ADD(-S). Projection 2D [7] uses the pose parameters to project the 3D model into the image plane and evaluates the result using the pixel difference between the predicted image and ground-truth image. In our experiment section, when the pixel difference is less than 5 px, the pose estimation is considered successful. For the evaluation of YCB-V dataset, we also calculate the AUC (area under curve) of ADD(-S) and the AUC (area under curve) of ADD-S by using different thresholds. The ADD-S metric employs the symmetric metric across all objects and the ADD(-S) only employs the symmetric metric for symmetric objects. Furthermore, we provide the evaluation results under BOP settings [38] on the LM-O dataset. The metric includes $AR_{VSD}$, $AR_{MSSD}$, and $AR_{MSPD}$. The Average Recall (AR) score is also presented, which is obtained by taking the average of $AR_{VSD}$, $AR_{MSSD}$, and $AR_{MSPD}$.

### 4.4. Comparison to the state of the art approaches

**LM datasets.** As shown in Table 1, our results outperform state-of-the-art methods on the LM dataset. LM is a dataset with slight occlusions, and our algorithm performs best among algorithms without refinement. Moreover, the proposed model even achieves higher results than algorithms with refinement. Compared with SOTA approaches, we have the best results for 7 out of 13 objects. Therefore, our algorithm can adapt well to slightly occluded scenes. In addition, Table 2 is presented that compares the performance of the proposed method with SOTA methods by using four evaluation metrics includes 2 cm 2°, 5 cm 5°, ADD(-S)-0.02d and Projection 2D 5px. Overall, the proposed method has the best performance than all other methods. Furthermore, under the more stringent metric 2 cm 2° and ADD(-S)-0.02d, the proposed method consistently outperforms the competing SOTA methods with 80.5 and 50.65, which validates the effectiveness of the proposed approach. The excellent performance under various metrics fully shows that the proposed OA-Pose can adapt well to slightly occlusion scenes.

**Table 2**
Results on LM referring to the n cm, n°, ADD(-S)-0.02d and Proj.2D 5px. metrics.

| Methods | 2 cm, 2° | 5 cm, 5° | ADD(-S)-0.02d | Proj.2D 5px. |
|---|---|---|---|---|
| DeepIM [7] | – | 85.2 | 30.9 | 97.5 |
| CDPN [13] | 59.3 | 94.3 | – | 98.1 |
| GDR-Net [16] | 62.1 | 95.6 | 35.5 | 98.72 |
| SO-Pose [18] | 76.9 | 98.5 | 45.9 | – |
| RNNPose [17] | – | – | 50.39 | – |
| Ours | **80.5** | **99.3** | **50.65** | **99.62** |

**LM-O datasets.** The LM-O datasets are collected in severe occlusion scenarios. We compare our method with other state-of-the-art methods on this dataset using ADD(-S) as metric. From Table 3, it can be seen that our method can obtain competitive results, even better than refinement-based methods. Therefore, it can be demonstrated from the results that our model has great potential for robot application in occlusion scenarios. Fig. 5 shows the visualization of pose prediction in LM-O dataset. The RGB image is used as input to first predict the 2D–3D correspondence at pixel level, and then calculate the pose parameters of the object. It can be seen that the pose of the occluded objects(*Can*, *Glue* and *Holepuncher*) can be well estimated. Fig. 6 provides a visual comparison of occluded object pose estimation between the proposed method and SOTA methods. Contrasting with two state-of-the-art methods that failed to accurately estimate the occluded objects' pose, the proposed method successfully determines the pose parameters. This demonstrates the superior performance of our approach under challenging conditions of object occlusion.

**YCB-V datasets.** Table 5 shows the evaluation results of the proposed method on the YCB-V dataset. We perform better than all other methods in AUC of ADD-S and AUC of ADD(-S), with 91.5 and 84.1 outperforming the second best method with 90.9 and 83.9. For ADD(-S), the proposed method outperforms the comparison methods without


**Fig. 7.** Visual comparison of pose estimation in occlusion scenes on YCB-V dataset.

**Table 3**
Results on LM-O referring to the Average Recall (%) of ADD(-S) metric.

| Object | w/o Refinement | | | | | | | w/ Refinement | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Single-Stage [22] | HybridPose [26] | ER-Pose [14] | BDR6D [12] | GDR-Net [16] | SO-Pose [18] | Ours | DeepIM [7] | RePose [23] | RNNPose [17] |
| Ape | 19.2 | 20.9 | 25.9 | 32.5 | 46.8 | 48.4 | 50.77 | **59.2** | 31.1 | 37.18 |
| Can | 65.1 | 75.3 | 72.1 | 81.6 | 90.8 | 85.8 | **92.29** | 63.5 | 80.0 | 88.07 |
| Cat | 18.9 | 24.9 | 25.3 | 37.1 | **40.5** | 32.7 | 38.08 | 26.2 | 25.6 | 29.15 |
| Driller | 69.0 | 70.2 | 72.9 | 75.7 | 82.6 | 77.4 | 82.95 | 55.6 | 73.1 | **88.14** |
| Duck | 25.3 | 27.9 | 35.8 | 38.4 | 46.9 | 48.9 | 48.68 | 52.4 | 43.0 | **49.17** |
| Eggbox | 52.0 | 52.4 | 48.7 | 60.2 | 54.2 | 52.4 | 53.20 | 63.0 | 51.7 | **66.98** |
| Glue | 51.4 | 53.8 | 58.8 | 52.9 | 75.8 | 78.3 | **79.13** | 71.7 | 54.3 | 63.79 |
| Holepuncher | 45.6 | 54.2 | 47.4 | 75.3 | 60.1 | 75.3 | **76.61** | 52.5 | 53.6 | 62.76 |
| Mean | 43.3 | 47.5 | 48.3 | 56.7 | 62.2 | 62.3 | **65.21** | 55.5 | 51.6 | 60.65 |

**Table 4**
Results on LM-O referring to the BOP metric. The best results are in bold and the second best results are in italics.

| Method | $AR_{VSD}$ | $AR_{MSSD}$ | $AR_{MSPD}$ | Mean $AR$ |
|---|---|---|---|---|
| PVNet [10] | 0.428 | 0.543 | 0.754 | 0.575 |
| CDPN [13] | 0.445 | **0.612** | 0.815 | 0.624 |
| GDR-Net [16] | **0.475** | 0.597 | 0.809 | *0.627* |
| SO-Pose [18] | 0.442 | 0.581 | *0.817* | 0.613 |
| ER-Pose [14] | 0.448 | 0.604 | 0.796 | 0.616 |
| Ours | *0.462* | *0.604* | **0.825** | **0.63** |

refinement, and has competitive results with RNNPose [17] that needs refinement. As shown in Fig. 7, for the YCB-V dataset, we select the occluded cracker box, the occluded meat can, and the occluded power driller as experiment objects. Compared with the SOTA method, even if the visible part of the object is missing, the proposed method can predict the pose parameters of the occluded object well.

**BOP Metrics.** Table 4 reports the experiment results on the LM-O dataset under BOP metric. We only use pbr data for training to achieve fair comparison. The proposed OA-Pose has an optimal $AR$ value of 0.63. At the same time, OA-Pose is also optimal in $AR_{MSPD}$, and reaches sub-optimal in $AR_{VSD}$ and $AR_{MSSD}$.

### 4.5. Application: Grasping experiments

Grasping object is the basic skill of robot [39], and we conduct the grasping experiment with the support of OA-Pose in the simulated and real-world robot. The Robotnik robot is used to verify the feasibility and performance of the proposed algorithm. As shown in Fig. 9, the grasping process is as follows: based on the robot vision system, the object pose in camera coordinate system is estimated by using OA-Pose. Then, the robot calculates the grasping trajectory, and the gripper moves to the target object position. Finally, the experiment is successful if the object is grasped and does not fall from the gripper.

For the experimental configuration, *power drill* and *can* from the YCB-V dataset [6] as experimental objects. The occlusion degree of the power drill and the can is from slight occlusion to extreme occlusion. Fig. 8 shows the pose estimation results under different occlusion degrees. Table 6 compares the grasping success rate of robots applying GDR-Net [16] and OA-Pose for a can. We conduct 30 experiments

**Table 5**
Results of YCB-V referring to the ADD(-S), AUC of ADD(-S) and AUC of ADD-S metrics.

| Metric | w/o Refinement | | | | | | w/ Refinement | |
|---|---|---|---|---|---|---|---|---|
| | PVNet [10] | GDR-Net [16] | SO-Pose [18] | Aing et al. [15] | PoET [9] | Ours | RePose [23] | RNNPose [17] |
| ADD(-S) | – | 49.1 | 56.8 | 51.8 | – | 60.5 | 60.3 | **66.4** |
| AUC of ADD(-S) | 73.4 | 80.2 | 83.9 | – | 80.8 | **84.1** | 80.8 | 83.1 |
| AUC of ADD-S | – | 89.1 | 90.9 | – | 87.1 | **91.5** | 86.7 | – |

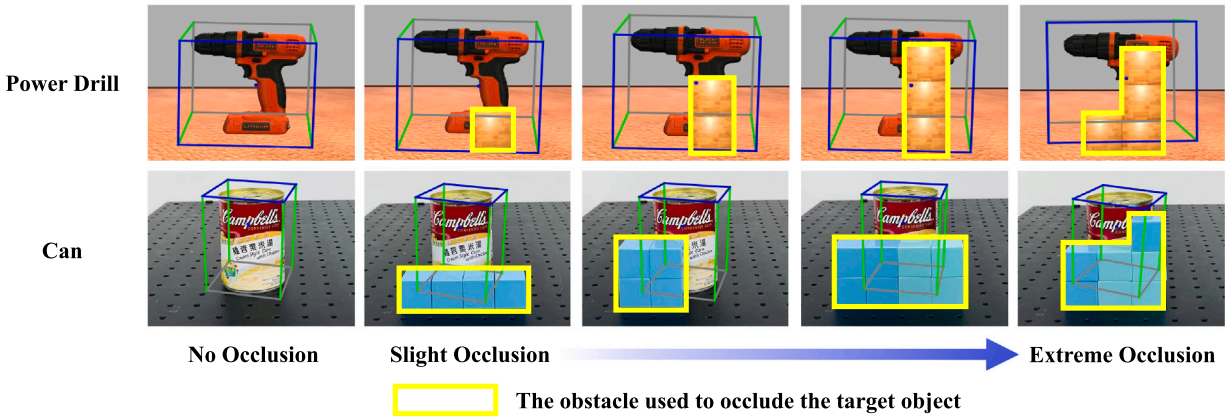


**Fig. 8.** Pose estimation experiments on different occlusion conditions. We provide the results of object poses (using the proposed model) under different occlusion degrees. For the *powerdrill*, we cover its handle with wooden blocks. Also, we use cubes to cover the *can*'s body.

**Table 6**
Comparison of robot grasping experiments based on pose estimation algorithms for a can.

| Method | Occlusion rate | Total | Success | Success rate |
|---|---|---|---|---|
| GDR-Net [16] | 0% is occluded | 30 | 30 | 100% |
| | 25% is occluded | 30 | 28 | 93.3% |
| | 50% is occluded | 30 | 25 | 83.3% |
| | 60% is occluded | 30 | 19 | 63.3% |
| SO-Pose [18] | 0% is occluded | 30 | 30 | 100% |
| | 25% is occluded | 30 | 28 | 93.3% |
| | 50% is occluded | 30 | 26 | 86.7% |
| | 60% is occluded | 30 | 20 | 66.7% |
| Ours | 0% is occluded | 30 | 30 | 100% |
| | 25% is occluded | 30 | 29 | 96.7% |
| | 50% is occluded | 30 | 25 | 83.3% |
| | 60% is occluded | 30 | 22 | 73.3% |

respectively, as shown in Table 6, with the increase of occlusion degree, the recognition accuracy gradually decreases. This is because the occlusion leads to less and less effective information in the image, which eventually leads to a decrease in accuracy. Meanwhile, compared with GDR-Net [16] and SO-Pose [18], our method has advantages under the same degree of occlusion. When the degree of occlusion is large, the proposed method can still maintain a success rate of more than 70%, which proves that OA-Pose is more suitable for cluttered environments. Fig. 9 shows two examples of a robot successfully grasping an occluded can.

### 4.6. Ablation experiment

In this section, based on the LM-O dataset (real data), we conduct the ablation study on Occlusion-Aware Geometry Alignment Loss and different backbones.

- Effectiveness of Occlusion-Aware Geometry Alignment Loss. In Table 7, we show the effectiveness of the proposed $L_{mi-2D}$ and $L_{mc-3D}$ loss functions on the LM-O dataset (real data). $L_{mi-2D}$ and $L_{mc-3D}$ are removed respectively to observe their effect on the 6D pose parameters w.r.t ADD(-S), 5° 5 cm and 10° 10 cm. As shown in Table 7, without either of these two loss functions, the metric decreases, which verifies the importance of Occlusion-Aware Geometry Alignment Loss.
- We use ResNeSt-50 [33] and ResNet-34 [40] as the backbone for ablation studies. From Table 7, when the network backbone is converted from ResNet-34 [40] to ResNeSt-50 [33], the accuracy will be significantly improved. Meanwhile, both backbones achieve higher results than GDR-Net [16]. Based on the same backbone (ResNeSt-50), the proposed method also performs better than SO-Pose [18] in occlusion scenes.

**Table 7**
Ablation study on LM-O (real). We report results for different loss functions and network backbones.

| | ADD(-S) | | 2 cm, 2° | 5 cm, 5° | 10 cm, 10° |
|---|---|---|---|---|---|
| | 0.05d | 0.1d | | | |
| GDR-Net [16] | 13.44 | 36.77 | 2.60 | 25.89 | 58.43 |
| SO-Pose [18] | 16.84 | 42.4 | 3.24 | 30.36 | 63.45 |
| Ours | **19.42** | **45.02** | **4.06** | **36.46** | **66.15** |
| w/o $L_{mi-2D}$ and $L_{mc-3D}$ | 16.06 | 41.65 | 3.27 | 31.18 | 63.26 |
| w/o $L_{mi-2D}$ | 18.15 | 43.01 | 3.97 | 34.28 | 63.51 |
| w/o $L_{mc-3D}$ | 17.74 | 42.07 | 3.37 | 34.05 | 63.24 |
| Ours (ResNet-34) | 16.35 | 39.32 | 3.37 | 29.49 | 61.04 |

### 4.7. Runtime analysis

Based on 640 × 480 images as input, on a computer with NVIDIA 2080 GPU, combined with Faster RCNN detector [31], we tested on the LM dataset and LM-O dataset. For single object, the proposed algorithm processing speed can reach 38 FPS. For all eight objects, it can reach 21 FPS.

### 4.8. Discussion

The proposed model provides an effective method for pose estimation of occluded objects. However, there are still some issues that urgently require more extensive and in-depth research in the future.

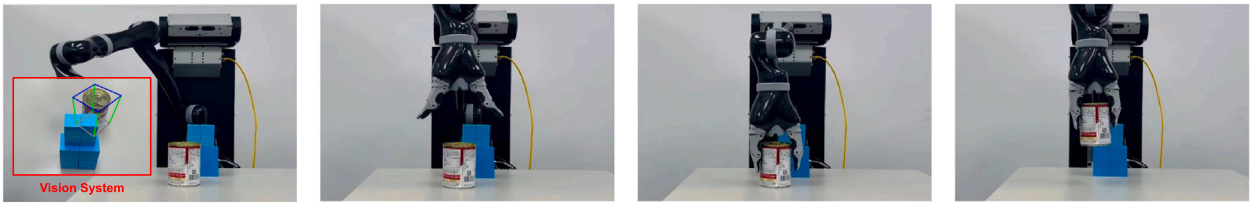


**Fig. 9.** The example of can is successfully grasped with the support of OA-Pose.

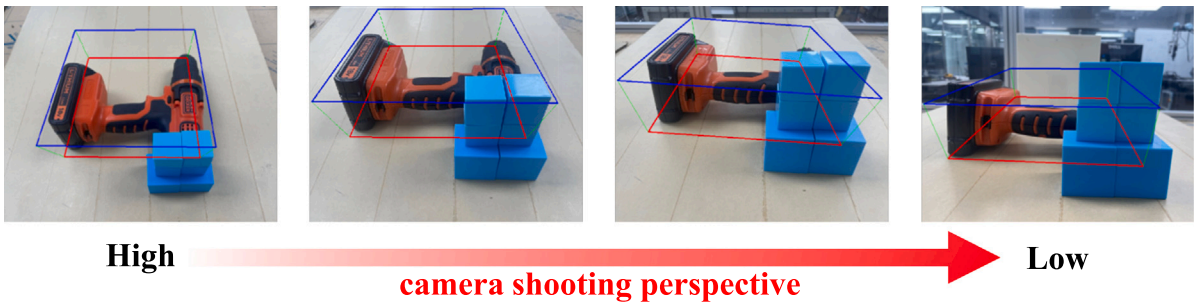


**Fig. 10.** The test results of pose estimation at different camera shooting perspective.

Firstly, we analyze the experimental results of the LM-O dataset in Table 3. The categories with lower accuracy are ape, duck and cat. These three types of objects are textureless, which contain too few feature points in RGB images. In occlusion scenes, since parts of the object are occluded by other objects, the visibility of feature points is further reduced, resulting in the algorithm being unable to obtain sufficient information for accurate pose estimation. To overcome this limitation, future research will explore the use of other types of features, such as edges, geometry, or depth information, to enhance pose estimation capabilities for occluded textureless objects.

Secondly, we study the impact of different camera shooting perspective on pose estimation. Fig. 10 is the test results of pose estimation at different camera shooting perspective. When the shooting perspective is too low, the results of pose estimation will be biased. This is because the performance of the algorithm, in addition to its own structure, also depends on the diversity and coverage of the training dataset. If the camera shooting perspective in the training dataset are limited, the algorithm may have difficulty processing objects at the lack of shooting perspective, resulting in reduced accuracy in pose estimation. In the future, we will consider more efficient dataset extension methods, such as efficient dataset production pipelines or augmenting existing datasets with synthetic data. Datasets based on multiple shooting perspectives and scene conditions ensure that the algorithm can be fully trained in various perspectives and scenes.

## 5. Conclusions

In this work, we propose a novel monocular 6D object pose estimation algorithm OA-Pose for occluded objects in clutter scenes with limited information provided. This work overcomes the difficulties for object perception in arbitrary illumination conditions and occluded scenarios, and thus enables intelligent robot manipulations in unstructured industrial practice environments. Unlike keypoint-based methods, the pixel-level 2D–3D correspondence is established in this work, which is dense and contains more accurate object information. The Occlusion-aware Geometry Alignment Module (OGA) and occlusion-aware loss functions are novelly proposed to drive the invisible correspondence toward the occluded parts of objects consistent across the 2D and 3D spaces. Extensive experiments demonstrate that our method outperforms the state-of-the-art public datasets, including LM, LM-O, and YCB-V. Moreover, the application experiments for robot grasping further verify that the proposed method can successfully conduct complex object manipulation tasks in unstructured real scenes.

In the future, in addition to the proposed model for object pose estimation in occlusion scenes, we plan to study the following two aspects further. The first is to address the challenges of lacking semantic features in 6-DoF pose estimation of texture-less objects under occluded scenarios. Secondly, it is essential to overcome the limitation in existing public datasets with insufficient camera shooting perspectives by proposing a comprehensive dataset production pipeline to enhance the robustness of pose estimation studies.

## CRediT authorship contribution statement

**Jikun Wang:** Methodology, Validation, Writing – original draft. **Luqing Luo:** Validation. **Weixiang Liang:** Data curation, Visualization. **Zhi-Xin Yang:** Conceptualization, Funding acquisition, Methodology, Writing – review & editing.

## Declaration of competing interest

All authors disclosed no relevant relationships and no conflict of interest.

## Data availability

Data will be made available on request.

## Acknowledgments

This work was funded in part by the Science and Technology Development Fund, Macau SAR (Grant no. 0075/2023/AMJ, 0003/2023/RIB1, 001/2024/SKL), in part by the Guangdong Science and Technology Department, China (Grant no. 2020B1515130001), in part by the Zhuhai Science and Technology Innovation Bureau, China (Grant no. ZH2220004002524), in part by the International Science and Technology project of Guangzhou Development District, China (Grant no. 2022GH09), in part by Zhuhai UM Research Institute, China (Grant no. HF-011-2021), and in part by the University of Macau (Grant No. : MYRG2022-00059-FST, MYRG-GRG2023-00237-FST-UMDF).

## References


[1] Rui-Dong Xi, Xiao Xiao, Tie-Nan Ma, Zhi-Xin Yang, Adaptive sliding mode disturbance observer based robust control for robot manipulators towards assembly assistance, IEEE Robot. Autom. Lett. 7 (3) (2022) 6139–6146.

[2] Jian Liu, Wei Sun, Chongpei Liu, Xing Zhang, Qiang Fu, Robotic continuous grasping system by shape transformer-guided multi-object category-level 6D pose estimation, IEEE Trans. Ind. Inform. (2023).

[3] Jikun Wang, Weixiang Liang, Jiangang Yang, Shizheng Wang, Zhi-Xin Yang, An adaptive image enhancement approach for safety monitoring robot under insufficient illumination condition, Comput. Ind. 147 (2023) 103862.

[4] Chenrui Wu, Long Chen, Shenglong Wang, Han Yang, Junjie Jiang, Geometric-aware dense matching network for 6D pose estimation of objects from RGB-D images, Pattern Recognit. (2023) 109293.

[5] Di Wang, Lulu Tang, Xu Wang, Luqing Luo, Zhi-Xin Yang, Improving deep learning on point cloud by maximizing mutual information across layers, Pattern Recognit. 131 (2022) 108892.

[6] Yu Xiang, Tanner Schmidt, Venkatraman Narayanan, Dieter Fox, Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes, 2017, arXiv preprint arXiv:1711.00199.

[7] Yi Li, Gu Wang, Xiangyang Ji, Yu Xiang, Dieter Fox, Deepim: Deep iterative matching for 6d pose estimation, in: Proceedings of the European Conference on Computer Vision, ECCV, 2018, pp. 683–698.

[8] Fabian Manhardt, Diego Martin Arroyo, Christian Rupprecht, Benjamin Busam, Tolga Birdal, Nassir Navab, Federico Tombari, Explaining the ambiguity of object detection and 6d pose from visual data, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 6841–6850.

[9] Thomas Georg Jantos, Mohamed Amin Hamdad, Wolfgang Granig, Stephan Weiss, Jan Steinbrener, PoET: pose estimation transformer for single-view, multi-object 6D pose estimation, in: Conference on Robot Learning, PMLR, 2023, pp. 1060–1070.

[10] Sida Peng, Yuan Liu, Qixing Huang, Xiaowei Zhou, Hujun Bao, Pvnet: Pixel-wise voting network for 6dof pose estimation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 4561–4570.

[11] Bugra Tekin, Sudipta N. Sinha, Pascal Fua, Real-time seamless single shot 6d object pose prediction, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 292–301.

[12] Penglei Liu, Qieshi Zhang, Jun Cheng, Bdr6d: Bidirectional deep residual fusion network for 6d pose estimation, IEEE Trans. Autom. Sci. Eng. (2023).

[13] Zhigang Li, Gu Wang, Xiangyang Ji, Cdpn: Coordinates-based disentangled pose network for real-time rgb-based 6-dof object pose estimation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 7678–7687.

[14] Xu Yang, Kunbo Li, Jinge Wang, Xiumin Fan, ER-Pose: Learning edge representation for 6D pose estimation of texture-less objects, Neurocomputing 515 (2023) 13–25.

[15] Lee Aing, Wen-Nung Lie, Guo-Shiang Lin, Faster and finer pose estimation for multiple instance objects in a single RGB image, Image Vis. Comput. 130 (2023) 104618.

[16] Gu Wang, Fabian Manhardt, Federico Tombari, Xiangyang Ji, GDR-net: Geometry-guided direct regression network for monocular 6D object pose estimation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 16611–16621.

[17] Yan Xu, Kwan-Yee Lin, Guofeng Zhang, Xiaogang Wang, Hongsheng Li, RNNPose: Recurrent 6-DoF object pose refinement with robust correspondence field estimation and pose optimization, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 14880–14890.

[18] Yan Di, Fabian Manhardt, Gu Wang, Xiangyang Ji, Nassir Navab, Federico Tombari, SO-pose: Exploiting self-occlusion for direct 6D pose estimation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, ICCV, 2021, pp. 12396–12405.

[19] Gu Wang, Fabian Manhardt, Xingyu Liu, Xiangyang Ji, Federico Tombari, Occlusion-aware self-supervised monocular 6D object pose estimation, IEEE Trans. Pattern Anal. Mach. Intell. (2021).

[20] Jian Liu, Wei Sun, Chongpei Liu, Xing Zhang, Shimeng Fan, Wei Wu, HFF6D: Hierarchical feature fusion network for robust 6D object pose tracking, IEEE Trans. Circuits Syst. Video Technol. 32 (11) (2022) 7719–7731.

[21] Jichun Wang, Lemiao Qiu, Guodong Yi, Shuyou Zhang, Yang Wang, Multiple geometry representations for 6D object pose estimation in occluded or truncated scenes, Pattern Recognit. 132 (2022) 108903.

[22] Yinlin Hu, Pascal Fua, Wei Wang, Mathieu Salzmann, Single-stage 6d object pose estimation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 2930–2939.

[23] Shun Iwase, Xingyu Liu, Rawal Khirodkar, Rio Yokota, Kris M. Kitani, Repose: Real-time iterative rendering and refinement for 6d object pose estimation, 2021, p. 6, arXiv preprint arXiv:2104.00633. 1 (2).

[24] Yannick Bukschat, Marcus Vetter, EfficientPose: An efficient, accurate and scalable end-to-end 6D multi object pose estimation approach, 2020, ArXiv, abs/2011.04307.

[25] Torsten Sattler, Qunjie Zhou, Marc Pollefeys, Laura Leal-Taixe, Understanding the limitations of cnn-based absolute camera pose regression, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 3302–3312.

[26] Chen Song, Jiaru Song, Qixing Huang, Hybridpose: 6d object pose estimation under hybrid representations, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 431–440.

[27] Junjie Jiang, Zaixing He, Xinyue Zhao, Shuyou Zhang, Chenrui Wu, Yang Wang, MLFNet: Monocular lifting fusion network for 6DoF texture-less object pose estimation, Neurocomputing 504 (2022) 16–29.

[28] Sergey Zakharov, Ivan Shugurov, Slobodan Ilic, Dpod: 6d pose object detector and refiner, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 1941–1950.

[29] Kiru Park, Timothy Patten, Markus Vincze, Pix2pose: Pixel-wise coordinate regression of objects for 6d pose estimation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 7668–7677.

[30] Eric Brachmann, Carsten Rother, Neural-guided RANSAC: Learning where to sample model hypotheses, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 4322–4331.

[31] Shaoqing Ren, Kaiming He, Ross Girshick, Jian Sun, Faster r-cnn: Towards real-time object detection with region proposal networks, Adv. Neural Inf. Process. Syst. 28 (2015).

[32] Yann Labbé, Justin Carpentier, Mathieu Aubry, Josef Sivic, Cosypose: Consistent multi-view multi-object 6d pose estimation, in: European Conference on Computer Vision, Springer, 2020, pp. 574–591.

[33] Hang Zhang, Chongruo Wu, Zhongyue Zhang, Yi Zhu, Haibin Lin, Zhi Zhang, Yue Sun, Tong He, Jonas Mueller, R. Manmatha, et al., Resnest: Split-attention networks, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 2736–2746.

[34] Stefan Hinterstoisser, Vincent Lepetit, Slobodan Ilic, Stefan Holzer, Gary Bradski, Kurt Konolige, Nassir Navab, Model based training, detection and pose estimation of texture-less 3d objects in heavily cluttered scenes, in: Asian Conference on Computer Vision, Springer, 2012, pp. 548–562.

[35] Eric Brachmann, Alexander Krull, Frank Michel, Stefan Gumhold, Jamie Shotton, Carsten Rother, Learning 6d object pose estimation using 3d object coordinates, in: European Conference on Computer Vision, Springer, 2014, pp. 536–551.

[36] Tomáš Hodaň, Martin Sundermeyer, Bertram Drost, Yann Labbé, Eric Brachmann, Frank Michel, Carsten Rother, Jiří Matas, BOP challenge 2020 on 6D object localization, in: European Conference on Computer Vision, Springer, 2020, pp. 577–594.

[37] Jamie Shotton, Ben Glocker, Christopher Zach, Shahram Izadi, Antonio Criminisi, Andrew Fitzgibbon, Scene coordinate regression forests for camera relocalization in RGB-D images, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 2930–2937.

[38] 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: Proceedings of the European Conference on Computer Vision, ECCV, 2018, pp. 19–34.

[39] Patricio Loncomilla, Javier Ruiz-del Solar, Luz Martínez, Object recognition using local invariant features for robotic applications: A survey, Pattern Recognit. 60 (2016) 499–514.

[40] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.



**Jikun Wang** received the M.S. degree and B.S. degree in China University of Geosciences, Beijing. He is currently an Ph.D. student in State Key Laboratory of Internet of Things for Smart City and Department of Electromechanical Engineering from the University of Macau. His research interests include 6D pose estimation and Robotic manipulation.

**Luqing Luo** received the Ph.D. degree in State Key Laboratory of Internet of Things for Smart City and Department of Electromechanical Engineering from the University of Macau, and the M.S. degree in Mechanical and Aerospace Engineering from North Carolina State University, USA. She is currently an assistant researcher from the Institute of Microelectronics of the Chinese Academy of Sciences. Her current research interests include integrated circuit-based computer vision, especially for 3D data analysis and semantic understanding.

**Weixiang Liang** received the B.S. degree in Shantou University, Shantou and M.S. degree in University of Macau, Macau. He is currently an Ph.D. student in State Key Laboratory of Internet of Things for Smart City and Department of Electromechanical Engineering from the University of Macau. His research interests include Deep Reinforcement Learning and Robotic manipulation.

**Zhi-Xin Yang** obtained the B.Eng. in mechanical engineering from the Huazhong University of Science and Technology, and the Ph.D. in industrial engineering and

engineering management from the Hong Kong University of Science and Technology, China. He is currently an Associate Professor with the State Key Laboratory of Internet of Things for Smart City and the Department of Electromechanical Engineering, University of Macau, Macau, China. His research interests include prognostic health monitoring of electromechanical systems, and computer vision based robotic systems.