2023 IEEE International Conference on Robotics and Automation (ICRA 2023)
May 29 - June 2, 2023. London, UK

RFFCE: Residual Feature Fusion and Confidence Evaluation Network for 6DoF Pose Estimation

Qiwei Meng^{1,2}, Shanshan Ji^{1,2}, Shiqiang Zhu^{1,2}, Tianlei Jin^{1,2}, Te Li^{1,2}, Jason Gu^{3} and Wei Song^{1,2}

2023 IEEE International Conference on Robotics and Automation (ICRA) | 979-8-3503-2365-8/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICRA48891.2023.10160448

Abstract— **In this paper, we propose a novel RGBD-based object 6DoF pose estimation network - RFFCE. It is a two-stage method that firstly leverages deep neural networks for feature extraction and object points matching, and then the geometric principles are utilized for final pose computation. Our approach consists of three primary innovations: residual feature fusion for representative RGBD feature extraction; confidence evaluation and confidence-based paired points offsets regression for self-evaluation and self-optimization respectively. Their effectiveness is verified through an ablation study, and our RFFCE achieves the SOTA performance on LineMOD, Occlusion-LineMOD and YCB-Video datasets. Additionally, we also conduct a real-world object grasping experiment for visualization and qualitative evaluation of the RFFCE.**

I. Introduction

6DoF pose estimation aims to evaluate the 3D location and orientation of target objects [1]. Unlike traditional vision problems like 2D detection and segmentation [2], 6DoF pose estimation demonstrates more abundant information of objects in the real world, thus enabling further operation and manipulation [3], [4]. The applications of 6DoF pose estimation are extensive, crossing from 3D reconstruction [5], [6], augmented reality [7], [8], robotic manipulation [3], [9], [10], and autonomous driving [11]. However, despite the significance and necessity of this technology in various areas, academia and industry have yet to develop a recognized and generic method [12], [13].

Compared with 2D vision tasks, 3D sensing would be more sensitive to background noise, sensor quality, scene occlusion and so on [14], [15], [16], thereby being fairly challenging in applications [13]. Minor fluctuations in sensing data could probably result in large errors of predictions [17]. Therefore, one-stage approaches that simply apply geometric information [18] or neural networks [19], [20], [21] for modeling can hardly achieve satisfying performances [1], [17], [22]. For robust pose estimation, researchers recently start to focus on developing two-stage hybrid approaches for modeling. A typical implementation [12], [17], [23], [24] is that deep neural networks (DNNs) are trained to extract features and locate paired points between image and object mesh models, and geometric algorithms are then applied to calculate 6DoF pose based on predicted paired points. Such hybrid methods not only take advantage of DNNs in investigating complex nonlinear relationships between images and object mesh models [25], but also consider the geometric constraints [26], [27]. Though hybrid approaches are proven to be more effective and promote the development of pose estimation, we notice that their performances are still far from satisfying, especially under occluded and cluttered scenarios, which limit the wide application of 6DoF estimation [28], [29].

In order to improve the performance of pose estimation algorithms and enable reliable real-world applications, it is significant to fully leverage the RGBD information of target object and construct a closed-loop visual system. Accordingly, we propose the RFFCE with three primary innovations: residual feature fusion, confidence evaluation and confidence-based paired points offsets regression. Residual feature fusion can facilitate the extraction of RGBD features by avoiding gradient vanishing during network training. Confidence evaluation and offsets regression are designed for reliable engineering practices under unstructured scenarios. Acting as self-assessment and self-optimization, they could sound warnings and adaptively fine-tune the prediction results in case of severely cluttered background or heavily occluded objects. To evaluate their effectiveness, LineMOD, Occlusion-LineMOD and YCB-Video are utilized as benchmarks, and our RFFCE achieves the SOTA performance among them. Additionally, a real-world grasping experiment is conducted for visualization and qualitative evaluation. In summary, the major contributions of our work are as follows:

1. We develop the residual feature fusion module to extract representative RGBD features of target object, which is also embeddable for other relevant RGBD vision tasks.

2. We propose the confidence evaluation and confidence-based paired points offsets regression modules, and they could serve as self-assessment and self-optimization to jointly improve model performance and reliability.

3. To the best of our knowledge, we achieve the SOTA performance on LineMOD, Occlusion-LineMOD and YCB-Video benchmarks.

4. We conduct a real-world grasping experiment to qualitatively demonstrate the practicability of RFFCE in engineering practices.

II. Related Work

A. Pose Estimation using RGB Data

The RGB methods take a single RGB image as input to calculate the pose of target objects. Traditional

This research is supported by Key Research Project of Zhejiang Lab (No. G2021NB0AL03).

[1]Research Center for Intelligent Robotics, Research Institute of Interdisciplinary Innovation, Zhejiang Lab, Hangzhou 311100, China, {mengqw, jiss, zhusq, jtl, lite, weisong}@zhejianglab.com.

[2]Zhejiang Engineering Research Center for Intelligent Robotics, Hangzhou 311100, China.

[3]Department of Electrical and Computer Engineering, Dalhousie University, Halifax, NS Canada B3H 4R2, jason.gu@dal.ca.

Zhu, Gu and Song are corresponding authors.

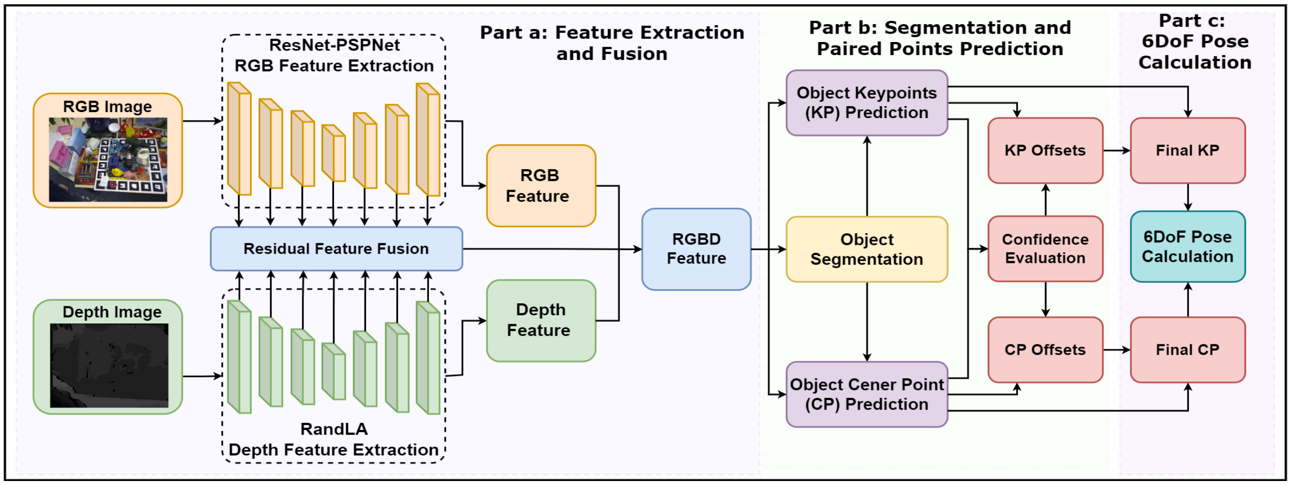
Part a: Feature Extraction and Fusion
Part b: Segmentation and Paired Points Prediction
Part c: 6DoF Pose Calculation
ResNet-PSPNet RGB Feature Extraction
RGB Image
RGB Feature
Residual Feature Fusion
Depth Image
Depth Feature
RandLA Depth Feature Extraction
RGBD Feature
Object Keypoints (KP) Prediction
Object Segmentation
Object Cener Point (CP) Prediction
KP Offsets
Confidence Evaluation
CP Offsets
Final KP
6DoF Pose Calculation
Final CP

Fig. 1: The Overview of RFFCE Model.

approaches [30], [31], [32] apply handcrafted features to align objects in the RGB image with their mesh models. Recently, with the development and popularity of deep learning, [33], [34] demonstrate the feasibility of applying end-to-end DNNs to directly compute 6DoF poses, while [35] respectively calculates 3D location and orientation using two-branch networks. However, these end-to-end methods are considered less robust under cluttered and occluded scenes [1], [36]. Therefore, [1] proposes a pixel-wise voting network to firstly locate 2D keypoints on the target image, and then apply the PnP solver for pose computation based on 2D - 3D keypoints correspondence. Generally, RGB methods would be effective for transparent objects and considerably faster [37], [34]. However, limited by sensing data available, their overall performances are not competitive enough.

B. *Pose Estimation using Point Cloud Data*

Recently, researchers gradually acknowledge the importance of 3D information in pose estimation, so lots of point cloud methods are coming forth [38], [39], [40]. Lidar and depth cameras are two commonly-used sensors for 3D information acquisition, their sensing data can be transformed into point cloud format for pose estimation [41]. [38], [39], [40] consider pose estimation to be a natural extension of finding the correspondence between point clouds and object mesh models, so they apply global optimal search and iterative closet point (ICP) algorithms for initial pose estimation and post-refinement respectively. [42], [43] emphasize the importance of initial pose and argue that a poor initial pose may render the post-refinement misleading and time-consuming. Therefore, they propose to apply DNNs for point-wise feature extraction [44], [45], and accordingly use these features for initial pose computation. The distinct characteristic of point cloud approaches is that 3D information is fully leveraged, which imposes geometric constraints on pose estimation.

C. *Pose Estimation using RGBD Data*

For this line of works, researchers apply RGBD data in various ways. [23], [46] use the RGB image mainly for object detection and interesting window selection, while the depth image is transformed into point clouds for feature extraction and pose calculation. With the advantages of deep learning, lots of studies have attempted to apply it for pose estimation, but it has been proven that end-to-end networks cannot obtain satisfying performances [21]. Therefore, preferable approaches [12], [17], [24], [47] are applying DNNs for RGBD feature extraction, and then leveraging extracted features to match paired points between the input image and object mesh models. Finally, the pose of target object can be calculated based on paired points correspondences. It is acknowledged that with abundant sensing data, RGBD methods are more competitive and promising. However, previous works fail to fully exploit the complementarity of RGB and depth information, so the representation of fused features is limited. In this paper, we propose residual feature fusion to align and mutually augment RGB and depth features, which significantly assists DNNs in feature learning.

III. Method

A. *Model Overview*

In this paper, we propose the RFFCE for 6DoF object pose estimation from a single RGBD image. As demonstrated in Fig.1, our RFFCE is a two-stage hybrid approach that combines the advantages of DNNs and geometric modeling, and it primarily includes three parts: a). feature extraction and fusion; b). segmentation and paired points prediction; c). 6DoF pose calculation. Given the scene RGBD image, ResNet-PSPNet and RandLA are trained to extract its RGB and depth features respectively, and meanwhile residual feature fusion module is applied to assist feature learning by leveraging the complementarity of RGBD features. Subsequently, taking extracted features as input, separate multi-layer perceptrons (MLPs) are designed to predict object segmentation, keypoints (KP), center points (CP) and estimation confidence. Finally, through combining prediction results with geometric principles, the pose of target object can be calculated through paired points correspondences. Details about major model components will be discussed below.

B. Residual Feature Fusion

The residual feature fusion module is proposed to extract representative object features in Model Part a. It is widely acknowledged that the fusion of RGB and depth features could be challenging due to their heterogeneity [1], [17], [24], so deep fused features are fairly sensitive to background noise, resulting in gradient vanishing and explosion during training [48]. Therefore, researchers have attempted various fusion schemes of RGBD features, like iterative dense fusion [24], bidirectional fusion [12], etc. Despite their contributions on improving overall model performance, the communications and mutual augmentation of RGBD features are still inadequate, so gradient-related problems remain unsolved and their models could probably fall into a local optimum at the end of training.

To avoid gradient vanishing and explosion, residual feature fusion module firstly aligns the initial layer-wise RGB and depth features in a unified space, where heterogeneous RGB and depth information can be mutually complemented and augmented. Subsequently, it applies an identity mapping structure to fuse aligned RGB and depth features for following network inference. As shown in Fig.1, the residual feature fusion module is deployed in each downsample/upsample layer of ResNet-PSPNet and RandLA. Taking the first and second downsample layers as an example, the implementation details of residual feature fusion are demonstrated in Fig.2. In the first layer, the pixel-wise RGB features $[n*c_{rgb1}*h_1*w_1]$ are mapped to 3D space and then concatenated with point-wise depth features $[n*c_{d1}*npts_1]$ for information sharing. Afterwards, the fused RGBD features $[n*(c_{d1}+c_{rgb1})*npts_1]$ are again decomposed into final RGB and depth features of layer1. Similarly, in the second layer, after alignment and fusion of RGBD features, the identity mapping brings together fused features of layer1 $[n*(c_{d1}+c_{rgb1})*npts_1]$ and layer2 $[n*(c_{d2}+c_{rgb2})*npts_2]$ for following feature decomposition, which ensures deep layers are still fed with representative and informative features, thereby relieving gradient vanishing and explosion.

Embedded in feature extraction phases, the residual fusion module effectively bridges corresponding layers of ResNet-PSPNet and RandLA, so RGB and depth information can fully communicate and mutually complement during the initial extraction. Additionally, the design of identity mapping and residual connection significantly reduce the risks of gradient-related problems. Therefore, more representative and pose-sensitive object features could be obtained for subsequent tasks.

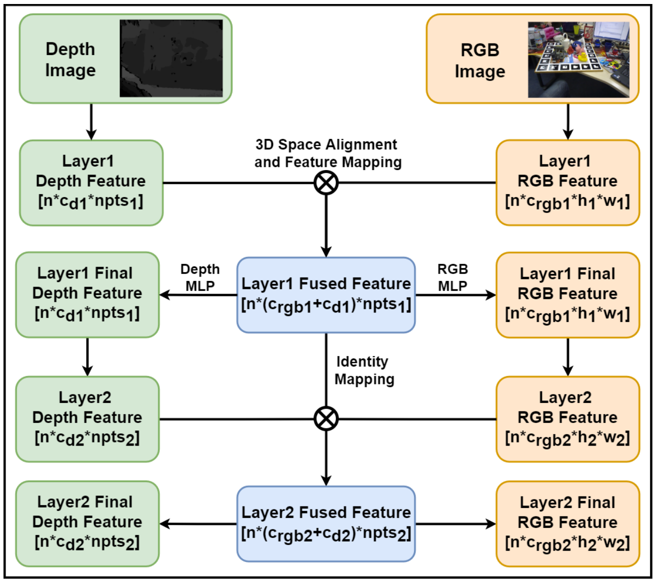
Depth Image
RGB Image
Layer1 Depth Feature $[n*c_{d1}*npts_1]$
3D Space Alignment and Feature Mapping
Layer1 RGB Feature $[n*c_{rgb1}*h_1*w_1]$
Layer1 Final Depth Feature $[n*c_{d1}*npts_1]$
Depth MLP
Layer1 Fused Feature $[n*(c_{rgb1}+c_{d1})*npts_1]$
RGB MLP
Layer1 Final RGB Feature $[n*c_{rgb1}*h_1*w_1]$
Identity Mapping
Layer2 Depth Feature $[n*c_{d2}*npts_2]$
Layer2 RGB Feature $[n*c_{rgb2}*h_2*w_2]$
Layer2 Final Depth Feature $[n*c_{d2}*npts_2]$
Layer2 Fused Feature $[n*(c_{rgb2}+c_{d2})*npts_2]$
Layer2 Final RGB Feature $[n*c_{rgb2}*h_2*w_2]$

Fig. 2: The Semantic Illustration of Residual Feature Fusion.

C. Confidence Evaluation and Offsets Regression

In Model Part b, separate MLPs are designed for object segmentation and KP/CP prediction based on extracted RGBD features. In this process, one commonly neglected issue in previous studies is the evaluation of prediction confidence [1], [12], [17], which performs a fairly important role in real-world applications like robotic manipulation, scene reconstruction, etc. Unfortunately, most studies attach greater importance to model performance improvement but overlook it, causing the gap between theoretical algorithms and engineering practices. Therefore, in this paper, we design the confidence evaluation module as a kind of self-assessment to improve the practicability of RFFCE.

Some DNN models [49], [50], [51] mainly depend on confidence intervals of historical data to conduct self-assessment. Nevertheless, we propose that for object pose estimation, the prediction confidence is more relevant with object feature quality and KP/CP reliability, rather than historical data distribution, so traditional statistical analysis methods are not appropriate. Accordingly, in the training phase of RFFCE, the prediction confidence is determined based on the degree of fit between predicted object KP/CP with ground truth, and its calculation is demonstrated in Eq.1 and 2.

$$\text{T\_Error} = \sum_{i=0}^{\text{select\_p}} \sum_{j=0}^{\text{paired\_p}} \left(\left|\Delta x_{ij}\right|+\left|\Delta y_{ij}\right|+\left|\Delta z_{ij}\right|\right) \tag{1}$$

$$\text{Conf} = \max\left\{0, \left(1-\frac{\text{T\_Error}}{\text{select\_p}\times\text{paired\_p}}\right)\right\} \tag{2}$$

where select_p is the number of valid voting points on target object; paired_p is the total number of KP and CP; $\triangle x_{ij}$, $\triangle y_{ij}$, and $\triangle z_{ij}$ are prediction errors of i$_{th}$ voting point on j$_{th}$ KP/CP in x, y and z directions respectively; T_Error is the absolute total prediction errors; Conf is the defined confidence score of target object.

After defining the prediction confidence score, an MLP is trained to regress this value with the initial KP/CP estimations as inputs. Hence, in the inference phase, we can obtain the prediction confidence along with estimation results, which provides crucial information for decision making and subsequent applications. Additionally, it is worth noting that we evaluate prediction confidence based on offsets in paired points, rather than 6DoF pose directly. This is mainly because the prediction errors in rotation (R) and translation (t) matrix can be sophisticated and difficult to compare [52], while offsets of KP and CP are more intuitive and measurable.

Apart from self-assessment, another important role of confidence scores is to regress KP and CP offsets for self-optimization. ICP is a widely applied algorithm for post-refinement, but it is also acknowledged to be time-consuming and even misleading if the initial estimation has large errors [42], [43]. Therefore, this paper applies learning-based approaches for KP/CP offsets prediction and post-refinement. Specifically, we design an MLP to take the initial KP/CP predictions and confidence score as inputs, and it can output KP/CP offsets, so the final KP/CP are the sum of initial predictions and offsets. Our offsets regression module includes two distinguishing characteristics, the first one is one-shot refinement to compute offsets in one shot rather than iteratively, thereby enhancing efficiency during inference. The second one is confidence-based offsets regression which utilizes object feature quality and initial prediction reliability for adaptive self-optimization, so the regressed offsets could be robust and effective under various application scenarios.

The introduction of confidence evaluation and offsets regression significantly improves the robustness of RFFCE in real-world applications. In case that the target object is heavily occluded, its prediction confidence will become correspondingly lower to warn the risks of manipulation based on current estimations, so camera can move to another viewpoint for better sensing data acquisition. In terms of slightly cluttered scenes or occluded objects, the self-optimization module could compute offsets of initial prediction and accordingly refine the results.

D. Clustering and Least-Squares Fitting

The Part c of RFFCE combines neural network predictions and geometric principles for 6DoF pose calculation. To this end, we first use the segmentation results to screen out image points on target object (denoted as select_p in Eq.1). Subsequently, the voted KP/CP from select_p are clustered through Meanshift algorithm to obtain final object KP/CP in camera coordinates system. Meanwhile, farthest point sampling (FPS) algorithm is applied on object mesh model for KP/CP generation in object coordinates system. Lastly, given two sets of KP/CP in camera and object coordinates system respectively, the least-squares algorithm is applied to calculate the best fitting rotation (R) and translation (t) matrix by minimizing the objective function in Eq.3:

$$\min_{\text{R,t}} \text{ Error} = \sum_{j=0}^{\text{paired\_p}} \left(\left\| \text{ point}_j - \left(R \times \text{point'}_j + t \right) \right\| \right) \qquad (3)$$

where paired_p is the total number of KP and CP; point_j and point'_j are 3D coordinates of j^{th} KP/CP in camera and object coordinates system respectively.

E. Loss Calculation

To supervise the training of RFFCE, we follow relevant works [1], [12], [17] to design the loss function and select hyperparameters. As shown in Eq.4, the loss function is composed of four parts: segmentation loss (L_{seg}); initial KP/CP loss ($L_{ini\_KP/CP}$), KP/CP offsets loss ($L_{off\_KP/CP}$) and confidence evaluation loss (L_{conf}).

$$L_{total} = \alpha L_{seg} + \beta L_{ini\_KP/CP} + \gamma L_{off\_KP/CP} + \delta L_{conf} \qquad (4)$$

where L_{seg} is focal loss [53] for point-wise segmentation; $L_{ini\_KP/CP}$ and $L_{off\_KP/CP}$ are both adaptive smooth L1 loss [54], they measure the spatial distance error between initial/refined predictions and ground truth; L_{conf} is mean absolute error loss (MAE) for confidence score regression.

IV. Experiments

A. Implementation Details

The RFFCE is coded based on PyTorch framework, and the hardware environments for model training are 24 Intel(R) Xeon(R) Gold 6248R CPU @ 3.00GHz and 8 Tesla V100S-PCIE-32GB GPU. The hyperparameter configurations applied in RFFCE mainly follow previous relevant studies [12], [17], [24] for a fair comparison: the number of KP and CP are 8 and 1 respectively, so paired_p is 9 in total; the α, β, γ, δ values in total loss calculation (Eq.5) are 2, 1, 0.1, 1 respectively. In the real-world grasping demo, RealSense D435i is utilized as ”eye-in-hand” sensor for RGBD data acquisition, and the robot arm is Aubo-i5.

B. Dataset

For the training and evaluation of RFFCE, three widely-used pose estimation benchmarksare selected.

LineMOD [18] consists of 13 objects, with approximately 1200 annotated RGBD images and 1 mesh model for each object. Following previous works [12], [17], we split this dataset into training and testing sets, and synthesis data are added into the training set since its number of images is fairly limited (around 200).

Occlusion-LineMOD [55] is an extension of LineMOD. It has similar data structure as LineMOD, but most objects are heavily occluded, making it more challenging. Following [1], [12], models are trained on LineMOD and this dataset is only utilized for testing.

YCB-Video [56] is a large RGBD dataset, which contains 92 videos for 21 objects. Similar to other learning-based methods [24], we apply 80 video sequences for model training and 2949 images from the rest 12 videos for testing.

C. Evaluation Metrics

To quantitatively assess the trained model, we mainly apply the average distance metrics (ADD and ADDS) for asymmetric and symmetric objects respectively. Their calculations are shown in Eq.5 and Eq.6:

$$ADD = \frac{1}{M} \times \sum_{i \in M}^{M} \left(\left\| (R_p \times i + t_p) - (R_g \times i + t_g) \right\| \right) \qquad (5)$$

$$ADDS = \frac{1}{M} \times \sum_{i \in M}^{M} \min_{j \in M} \left(\left\| (R_p \times i + t_p) - (R_g \times j + t_g) \right\| \right) \quad (6)$$

where M is the number of points on object model; R_p, t_p, R_g, t_g are predicted and ground truth pose matrix respectively.

TABLE I: Quantitative Evaluation on LineMOD Dataset. ADD(S)-0.1d is tabulated and symmetric objects are in bold.

Object	Model				
	DF[24]	G2L[23]	PVN3D[17]	FFB6D[12]	Ours
Ape	92.3	96.8	97.3	98.4	**99.14**
Benchvise	93.2	96.1	99.7	**100**	**100**
Camera	94.4	98.2	99.6	99.9	**100**
Can	93.1	98	99.5	99.8	**100**
Cat	96.5	99.2	99.8	99.9	**100**
Driller	87	99.8	99.3	**100**	**100**
Duck	92.3	97.7	98.2	98.4	**99.06**
Eggbox	99.8	**100**	99.8	**100**	**100**
Glue	**100**	**100**	**100**	**100**	**100**
Holepuncher	92.1	99	**99.9**	99.8	**99.9**
Iron	97	99.3	99.7	99.9	**100**
Lamp	95.3	99.5	99.8	99.9	**100**
Phone	92.8	98.9	99.5	99.7	**100**
MEAN	94.3	98.7	99.4	99.7	**99.85**

TABLE II: Quantitative Evaluation on Occlusion-LineMOD Dataset. ADD(S)-0.1d is tabulated and symmetric objects are in bold.

Object	Model				
	PVNet[1]	HPose[57]	PVN3D[17]	FFB6D[12]	Ours
Ape	15.8	20.9	33.9	47.2	**59.2**
Can	63.3	75.3	88.6	85.2	**96.1**
Cat	16.7	24.9	39.1	45.7	**47.9**
Driller	65.7	70.2	78.4	81.4	**95.4**
Duck	25.2	27.9	41.9	53.9	**62.7**
Eggbox	50.2	52.4	**80.9**	70.2	59.4
Glue	49.6	53.8	**68.1**	60.1	57.4
Holepuncher	39.7	54.2	74.7	85.9	**87.4**
MEAN	40.8	47.5	63.2	66.2	**70.7**

Additionally, the average distance errors are closely associated with the size of objects, so we apply ADD(S)-0.1d for a fair comparison of objects in LineMOD and Occlusion-LineMOD [1], [24], which is the percentage of points with ADD(S) less than 10% of the object′s diameter. As for YCB-Video dataset, in addition to ADD and ADDS, the AUC is proposed for evaluation following previous studies [12], [17], [24], which represents the area under ADD(S) accuracy threshold (0.1m). Specifically, ADDS-AUC and ADD(S)-AUC are applied, the first one calculates ADDS-AUC for both asymmetric and symmetric objects, while the latter one calculates ADD-AUC for asymmetric objects and ADDS-AUC for symmetric objects.

D. Quantitative Evaluations

TABLE I demonstrates the quantitative results of our RFFCE on LineMOD dataset. It is manifest that our model accomplishes the SOTA performance in average ADD(S)-0.1d, and for individual objects, the improvement is most significant in ape and duck.

For trained models from LineMOD dataset, their performances on Occlusion-LineMOD are shown in TABLE II. Compared with the previous SOTA approaches, our RFFCE improves the average ADD(S)-0.1d by 4.5%, suggesting its strong robustness against occlusion. Besides, we also notice that the improvement is particularly obvious for asymmetric objects, while for symmetric objects, the heavy occlusion in RGB images negatively influences their symmetry, thus lowering the performance.

Regarding the YCB-Video dataset, its evaluation results are summarized in TABLE III. We could observe that without computationally complex ICP for post-refinement, our model still achieves competitive performance on two primary evaluation metrics, mean ADDS-AUC and mean ADD(S)-AUC, again proving the superior structure design.

E. Ablation Study

This paper proposes three primary innovations, with residual feature fusion (RFF) and confidence-based paired points offsets regression (OR) directly contributing to 6DoF pose accuracy. To verify their effectiveness, an ablation study is conducted. TABLE IV demonstrates the ADD(S)-0.1d for representative objects in LineMOD and Occlusion-LineMOD. It is manifest that with stand-alone application of RFF or OR module, the baseline model improves obviously, and these two modules can also be combined to achieve the SOTA performance. Furthermore, TABLE V shows the mean ADDS-AUC and mean ADD(S)-AUC of 21 objects in YCB-Video, and we can notice that the RFF and OR modules can enhance model performance both separately and jointly. In addition, RFF and OR are both embeddable and compatible, so they can be adaptively generalized in other relevant works to improve feature representativity and robustness.

F. Results Visualization

For qualitative evaluation of proposed RFFCE, the visualization results on LineMOD, Occlusion-LineMOD and YCB-Video are shown in Fig.3 - Fig.5, where white annotations are predicted confidence scores for individual objects; points in various colors are the estimated 2D projection of objects model; and red and blue 3D bounding boxes represent the ground truth and estimated object poses respectively. It is apparent that RFFCE achieves satisfying performances, even with low prediction confidence, the final results are acceptable with the help of offsets regression module, suggesting the good robustness and broad applicability of our model.

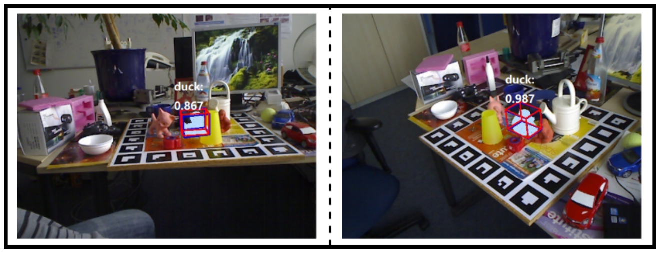

Fig. 3: Visualization on LineMOD.

In addition, we conduct a robotic grasping experiment to further verify the performance and practicability of RFFCE under real-world scenarios. Fig.6 visualizes the partial qualitative results during object grasping. It is manifest that our RFFCE can effectively estimate the 6DoF pose for randomly placed objects, and the estimation precision meets requirements of grasping and manipulation, even with lighting variations and cluttered scenes.

TABLE III: Quantitative Evaluation on YCB-Video Dataset. ADDS-AUC, ADD(S)-AUC (abbreviated as ADDS and ADD(S)) are tabulated and symmetric objects are in bold.

Object	Model (Without Iterative Refinement)									
	DF[24]		G2L[23]		PVN3D[17]		FFB6D[12]		Ours	
	ADDS	ADD(S)	ADDS	ADD(S)	ADDS	ADD(S)	ADDS	ADD(S)	ADDS	ADD(S)
002_master_chef_can	95.3	70.7	94	/	96	80.5	**96.3**	**80.6**	96.0	77.6
003_cracker_box	92.5	86.9	88.7	/	96.1	94.8	96.3	94.6	**96.4**	**94.9**
004_sugar_box	95.1	90.8	96	/	97.4	96.3	97.6	96.6	**97.8**	**96.9**
005_tomato_soup_can	93.8	84.7	86.4	/	**96.2**	88.5	95.6	**89.6**	**96.2**	87.4
006_mustard_bottle	95.8	90.9	95.9	/	97.5	96.2	97.8	97	**98.0**	**97.1**
007_tuna_fish_can	95.7	79.6	96	/	96	**89.3**	96.8	88.9	**97.1**	87.5
008_pudding_box	94.3	89.3	93.5	/	**97.1**	**95.7**	**97.1**	94.6	96.5	93.3
009_gelatin_box	97.2	95.8	96.8	/	97.7	96.1	**98.1**	**96.9**	97.9	96.4
010_potted_meat_can	89.3	79.6	86.2	/	93.3	88.6	94.7	88.1	**95.5**	**91.7**
011_banana	90	76.7	96.3	/	96.6	93.7	97.2	94.9	**97.7**	**96.0**
019_pitcher base	93.6	87.1	91.8	/	97.4	96.5	97.6	96.9	**97.9**	**97.3**
021_bleach_cleanser	94.4	87.5	92	/	96	93.2	**96.8**	**94.8**	96.6	94.1
024_bowl	86	86	86.7	/	90.2	90.2	**96.3**	**96.3**	95.2	95.2
025_mug	95.3	83.8	95.4	/	**97.6**	95.4	97.3	94.2	**97.6**	**95.6**
035_power_drill	92.1	83.7	95.2	/	96.7	95.1	97.2	95.9	**97.3**	**96.1**
036_wood_block	89.5	89.5	86.2	/	90.4	90.4	92.6	92.6	**94.0**	**94.0**
037_scissors	90.1	77.4	83.8	/	96.7	92.7	**97.7**	**95.7**	97.3	95.3
040_large_marker	95.1	89.1	96.8	/	96.7	**91.8**	96.6	89.1	**96.8**	**91.8**
051_large_clamp	71.5	71.5	94.4	/	93.6	93.6	96.8	96.8	**96.9**	**96.9**
052_extra_large_clamp	70.2	70.2	92.3	/	88.4	88.4	**96**	**96**	**96.0**	**96.0**
061_foam_brick	92.2	92.2	94.7	/	96.8	96.8	97.3	97.3	**98.0**	**98.0**
MEAN	91.2	82.9	92.4	/	95.5	91.8	96.6	92.7	**96.8**	**93.8**

TABLE IV: Ablation Study on LineMOD and Occlusion-LineMOD (abbreviated as Occ-LM) Dataset. ADD(S)-0.1d is tabulated.

RFF	OR	LineMOD				Occ-LM	
		Ape	Cam	Duck	Phone	Ape	Duck
		98.48	99.9	98.4	99.81	54.1	55.3
✓		99.05	**100**	98.97	**100**	58.8	56.4
	✓	98.67	**100**	98.59	**100**	57.6	57.2
✓	✓	**99.14**	**100**	**99.06**	**100**	**59.2**	**62.7**

TABLE V: Ablation Study on YCB-Video Dataset. Mean ADDS-AUC and ADD(S)-AUC are tabulated.

RFF	OR	YCB-Video	
		Mean ADDS-AUC	Mean ADD(S)-AUC
		96.2	92.8
✓		96.5	93.4
	✓	96.3	93.3
✓	✓	**96.8**	**93.8**

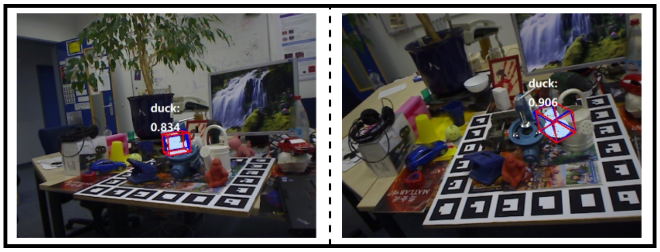

Fig. 4: Visualization on Occlusion-LineMOD.

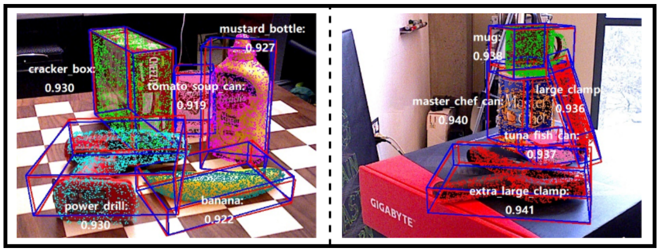

Fig. 5: Visualization on YCB-Video.

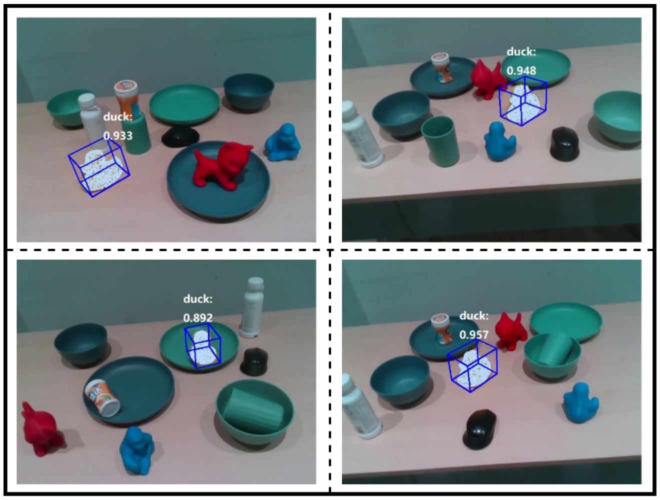

Fig. 6: Visualization for Robotic Grasping Demo.

V. CONCLUSION

To conclude, we propose a novel network RFFCE for object 6DoF pose estimation from a single RGBD image, which includes three primary innovations, residual feature fusion, confidence evaluation and confidence-based paired points offsets regression. The first one improves the representativity and sensitivity of extracted features, while the latter two not only enhance the model robustness against occlusion, but also fill the gap between pose estimation and engineering practices. With their implementation, our RFFCE outperforms previous SOTA approaches on three acknowledged pose estimation benchmarks, especially the Occlusion-LineMOD. Additionally, through a real-world experiment, the RFFCE is proven to be applicable for object grasping and manipulation, even under cluttered scenes.

References

[1] S. Peng, Y. Liu, Q. Huang, X. Zhou, and H. Bao, "Pvnet: Pixel-wise voting network for 6dof pose estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4561–4570, 2019.

[2] W. Zhang, W. Zhang, K. Liu, and J. Gu, "A feature descriptor based on local normalized difference for real-world texture classification," *IEEE Transactions on Multimedia*, vol. 20, no. 4, pp. 880–888, 2017.

[3] 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 *European Conference on Computer Vision (ECCV)*, pp. 205–220, Springer, 2016.

[4] S. Stevšić, S. Christen, and O. Hilliges, "Learning to assemble: Estimating 6d poses for robotic object-object manipulation," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 1159–1166, 2020.

[5] A. Dame, V. A. Prisacariu, C. Y. Ren, and I. Reid, "Dense reconstruction using 3d object shape priors," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1288–1295, 2013.

[6] D.-C. Hoang, T. Stoyanov, and A. J. Lilienthal, "Object-rpe: Dense 3d reconstruction and pose estimation with convolutional neural networks for warehouse robots," in *2019 European Conference on Mobile Robots (ECMR)*, pp. 1–6, IEEE, 2019.

[7] F. Tang, Y. Wu, X. Hou, and H. Ling, "3d mapping and 6d pose computation for real time augmented reality on cylindrical objects," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 9, pp. 2887–2899, 2019.

[8] Y. Su, J. Rambach, N. Minaskan, P. Lesur, A. Pagani, and D. Stricker, "Deep multi-state object pose estimation for augmented reality assembly," in *2019 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*, pp. 222–227, IEEE, 2019.

[9] T. Grundmann, R. Eidenberger, M. Schneider, M. Fiegert, and G. v. Wichert, "Robust high precision 6d pose determination in complex environments for robotic manipulation," in *Proc. Workshop Best Practice in 3D Perception and Modeling for Mobile Manipulation at the Int. Conf. Robotics and Automation (ICRA)*, pp. 1–6, 2010.

[10] X. Deng, Y. Xiang, A. Mousavian, C. Eppner, T. Bretl, and D. Fox, "Self-supervised 6d object pose estimation for robot manipulation," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 3665–3671, IEEE, 2020.

[11] Y. Liu, Y. Yixuan, and M. Liu, "Ground-aware monocular 3d object detection for autonomous driving," *IEEE Robotics and Automation Letters*, vol. 6, no. 2, pp. 919–926, 2021.

[12] Y. He, H. Huang, H. Fan, Q. Chen, and J. Sun, "Ffb6d: A full flow bidirectional fusion network for 6d pose estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3003–3013, 2021.

[13] Z. He, W. Feng, X. Zhao, and Y. Lv, "6d pose estimation of objects: Recent technologies and challenges," *Applied Sciences*, vol. 11, no. 1, p. 228, 2021.

[14] O. Hosseini Jafari, S. K. Mustikovela, K. Pertsch, E. Brachmann, and C. Rother, "ipose: instance-aware 6d pose estimation of partly occluded objects," in *Asian Conference on Computer Vision (ACCV)*, pp. 477–492, Springer, 2018.

[15] Y. Ma, D. Lin, B. Zhang, Q. Liu, and J. Gu, "A novel algorithm of image gaussian noise filtering based on pcnn time matrix," in *2007 IEEE International Conference on Signal Processing and Communications*, pp. 1499–1502, IEEE, 2007.

[16] A. Bais, R. Sablatnig, and J. Gu, "Single landmark based self-localization of mobile robots," in *The 3rd Canadian Conference on Computer and Robot Vision (CRV'06)*, pp. 67–67, IEEE, 2006.

[17] Y. He, W. Sun, H. Huang, J. Liu, H. Fan, and J. Sun, "Pvn3d: A deep point-wise 3d keypoints voting network for 6dof pose estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11632–11641, 2020.

[18] S. Hinterstoisser, V. Lepetit, S. Ilic, S. Holzer, G. Bradski, K. Konolige, and N. Navab, "Model based training, detection and pose estimation of texture-less 3d objects in heavily cluttered scenes," in *Asian Conference on Computer Vision (ACCV)*, pp. 548–562, Springer, 2012.

[19] J. Wu, B. Zhou, R. Russell, V. Kee, S. Wagner, M. Hebert, A. Torralba, and D. M. Johnson, "Real-time object pose estimation with pose interpreter networks," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 6798–6805, IEEE, 2018.

[20] Y. Hu, P. Fua, W. Wang, and M. Salzmann, "Single-stage 6d object pose estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2930–2939, 2020.

[21] V. Balntas, A. Doumanoglou, C. Sahin, J. Sock, R. Kouskouridas, and T.-K. Kim, "Pose guided rgbd feature learning for 3d object pose estimation," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 3856–3864, 2017.

[22] 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 *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 3828–3836, 2017.

[23] W. Chen, X. Jia, H. J. Chang, J. Duan, and A. Leonardis, "G2l-net: Global to local network for real-time 6d pose estimation with embedding vector features," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4233–4242, 2020.

[24] C. Wang, D. Xu, Y. Zhu, R. Martín-Martín, C. Lu, L. Fei-Fei, and S. Savarese, "Densefusion: 6d object pose estimation by iterative dense fusion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3343–3352, 2019.

[25] J. V. Tu, "Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes," *Journal of clinical epidemiology*, vol. 49, no. 11, pp. 1225–1231, 1996.

[26] Y. Wen, Y. Fang, J. Cai, K. Tung, and H. Cheng, "Gccn: Geometric constraint co-attention network for 6d object pose estimation," in *Proceedings of the 29th ACM International Conference on Multimedia*, pp. 2671–2679, 2021.

[27] C. Mitash, A. Boularias, and K. Bekris, "Robust 6d object pose estimation with stochastic congruent sets," *arXiv preprint arXiv:1805.06324*, 2018.

[28] A. Hietanen, J. Latokartano, A. Foi, R. Pieters, V. Kyrki, M. Lanz, and J.-K. Kämäräinen, "Benchmarking pose estimation for robot manipulation," *Robotics and Autonomous Systems*, vol. 143, p. 103810, 2021.

[29] A. Hietanen, J. Latokartano, A. Foi, R. Pieters, V. Kyrki, M. Lanz, and J.-K. Kämäräinen, "Object pose estimation in robotics revisited," *arXiv preprint arXiv:1906.02783*, 2019.

[30] M. Aubry, D. Maturana, A. A. Efros, B. C. Russell, and J. Sivic, "Seeing 3d chairs: exemplar part-based 2d-3d alignment using a large dataset of cad models," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3762–3769, 2014.

[31] F. Rothganger, S. Lazebnik, C. Schmid, and J. Ponce, "3d object modeling and recognition using local affine-invariant image descriptors and multi-view spatial constraints," *International journal of computer vision*, vol. 66, no. 3, pp. 231–259, 2006.

[32] A. Collet, M. Martinez, and S. S. Srinivasa, "The moped framework: Object recognition and pose estimation for manipulation," *The international journal of robotics research*, vol. 30, no. 10, pp. 1284–1306, 2011.

[33] Y. Xiang, T. Schmidt, V. Narayanan, and D. Fox, "Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes," *arXiv preprint arXiv:1711.00199*, 2017.

[34] T.-T. Do, M. Cai, T. Pham, and I. Reid, "Deep-6dpose: Recovering 6d object pose from a single rgb image," *arXiv preprint arXiv:1802.10367*, 2018.

[35] Y. Li, G. Wang, X. Ji, Y. Xiang, and D. Fox, "Deepim: Deep iterative matching for 6d pose estimation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 683–698, 2018.

[36] B. Tekin, S. N. Sinha, and P. Fua, "Real-time seamless single shot 6d object pose prediction," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 292–301, 2018.

[37] W. Kehl, F. Manhardt, F. Tombari, S. Ilic, and N. Navab, "Ssd-6d: Making rgb-based 3d detection and 6d pose estimation great again," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 1521–1529, 2017.

[38] J. P. S. do Monte Lima and V. Teichrieb, "An efficient global point cloud descriptor for object recognition and pose estimation," in *2016 29th SIBGRAPI conference on graphics, patterns and images (SIBGRAPI)*, pp. 56–63, IEEE, 2016.

[39] G. Izatt, H. Dai, and R. Tedrake, "Globally optimal object pose estimation in point clouds with mixed-integer programming," in *Robotics Research*, pp. 695–710, Springer, 2020.

[40] L. Liu, G. Zhao, and Y. Bo, "Point cloud based relative pose estimation of a satellite in close range," *Sensors*, vol. 16, no. 6, p. 824, 2016.

[41] R. B. Rusu and S. Cousins, "3d is here: Point cloud library (pcl)," in *2011 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1–4, IEEE, 2011.

[42] W. Tang, D. Zou, and P. Li, "Learning-based point cloud registration: A short review and evaluation," in *2021 2nd International Conference on Artificial Intelligence in Electronics Engineering*, pp. 27–34, 2021.

[43] L. Li, R. Wang, and X. Zhang, "A tutorial review on point cloud registrations: Principle, classification, comparison, and technology challenges," *Mathematical Problems in Engineering*, vol. 2021, 2021.

[44] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "Pointnet++: Deep hierarchical feature learning on point sets in a metric space," *Advances in neural information processing systems*, vol. 30, 2017.

[45] Q. Hu, B. Yang, L. Xie, S. Rosa, Y. Guo, Z. Wang, N. Trigoni, and A. Markham, "Randla-net: Efficient semantic segmentation of large-scale point clouds," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11108–11117, 2020.

[46] U. Asif, M. Bennamoun, and F. Sohel, "Real-time pose estimation of rigid objects using rgb-d imagery," in *2013 IEEE 8th Conference on Industrial Electronics and Applications (ICIEA)*, pp. 1692–1699, IEEE, 2013.

[47] T. Hodan, F. Michel, E. Brachmann, W. Kehl, A. GlentBuch, D. Kraft, B. Drost, J. Vidal, S. Ihrke, X. Zabulis, *et al.*, "Bop: Benchmark for 6d object pose estimation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 19–34, 2018.

[48] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

[49] B. Jiang, X. Zhang, and T. Cai, "Estimating the confidence interval for prediction errors of support vector machine classifiers," *The Journal of Machine Learning Research*, vol. 9, pp. 521–540, 2008.

[50] G. Chryssolouris, M. Lee, and A. Ramsey, "Confidence interval prediction for neural network models," *IEEE Transactions on neural networks*, vol. 7, no. 1, pp. 229–232, 1996.

[51] R. Dybowski, "Assigning confidence intervals to neural network predictions," in *Neural Computing Applications Forum (NCAF) Conference*, 1997.

[52] D. Q. Huynh, "Metrics for 3d rotations: Comparison and analysis," *Journal of Mathematical Imaging and Vision*, vol. 35, no. 2, pp. 155–164, 2009.

[53] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international Conference on Computer Vision (ICCV)*, pp. 2980–2988, 2017.

[54] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE international conference on computer vision*, pp. 1440–1448, 2015.

[55] E. Brachmann, A. Krull, F. Michel, S. Gumhold, J. Shotton, and C. Rother, "Learning 6d object pose estimation using 3d object coordinates," in *European conference on computer vision*, pp. 536–551, Springer, 2014.

[56] B. Calli, A. Singh, A. Walsman, S. Srinivasa, P. Abbeel, and A. M. Dollar, "The ycb object and model set: Towards common benchmarks for manipulation research," in *2015 international Conference on Advanced Robotics (ICAR)*, pp. 510–517, IEEE, 2015.

[57] C. Song, J. Song, and Q. Huang, "Hybridpose: 6d object pose estimation under hybrid representations," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 431–440, 2020.