Weak6D: Weakly Supervised 6D Pose Estimation With Iterative Annotation Resolver

Fengjun Mu, Rui Huang, *Member, IEEE*, Kecheng Shi, Xin Li, Jing Qiu, and Hong Cheng, *Senior Member, IEEE*

***Abstract*—6D object pose estimation is an essential task in vision-based robotic grasping and manipulation. Prior works always train models with a large number of pose annotated images, limiting the efficiency of model transfer between different scenarios. This letter presents an end-to-end model named *Weak6D*, which could be learned with unannotated RGB-D data. The core of the proposed approach is the novel optimizing method Iterative Annotation Resolver, which has the ability to directly utilize the captured RGB-D data through the training process. Furthermore, we employ a weak refinement loss to optimize the pose estimation network with refined object poses. We evaluated the proposed *Weak6D* in the YCB-Video dataset, and experimental results show our model achieved practical results without annotated data.**

***Index Terms*—Computer vision, iterative methods, object pose estimation, weakly-supervised learning.**

I. Introduction

6D OBJECT pose estimation has gained considerable interest in robotic grasping and manipulation [1], [2], especially in industrial-related applications [3], [4]. An ideal industrial solution of robotic grasping and manipulation is that the manipulator should deal with various objects. With the development of inexpensive RGB-D sensors, pose estimation methods with RGB-D data achieve more accuracy than methods with only RGB data, and allows us to perform large-scale capturing RGB-D data at a low cost. However, existing pose estimation methods always rely on a large number of high-cost 6D pose annotations, which is difficult to transfer the pose estimation model efficiently between different scenarios with different objects.

Traditional pose estimation methods generate templates or extract features from RGB-D data and find the best match using distance measures [5]–[7]. However, these methods are sensitive to occlusions, lighting variations, and cluttered environments. Recently deep neural networks achieved great successes in RGB-D-based pose estimation [8]–[10]. PoseCNN is one of the first end-to-end models to predict 6D object pose from RGB-D input [8]. In order to satisfy requirements of both accurate pose estimation and fast inference in real-time tasks, many dense feature-based methods with low-cost refinement process are presented [11]–[13]. This kind of deep learning-based method can achieve good results. However, these methods require 6D pose annotations as ground truth in the model training process, which limits the efficiency of model transfer between different application scenarios.

In order to estimate the 6D pose of different objects with less annotated data, refinement methods are employed to enhance the model performance. Iterative Closest Point (ICP) method is commonly used for point cloud registration to optimize the results of pose estimation process [8], [14]. In order to achieve better speed and performance of the refinement step, neural network-based refinement optimization methods are proposed [12], [13]. However, these methods still require the optimized model to generate high-precision pose estimation results. Therefore, these methods still relies on the basic model that needs to be trained with a large amount of annotated data.

In this letter, we propose a novel pose estimation method *Weak6D* as Fig. 1, which could be learned with unannotated RGB-D data. The core of the proposed approach is the novel optimizing method Iterative Annotation Resolver, which has the ability to utilize the captured RGB-D data for model training directly. Furthermore, we employ a weak refinement loss to optimize the pose estimation network with refined object poses. Compared to the existing methods [12], [13], *Weak6D* does not need high-cost 6D pose annotation of the training process, which greatly reduces the difficulty of model transfer in different industrial scenarios.

The proposed model is evaluated in a popular benchmark *YCB-Video* [8] for 6D object pose estimation. We carry out two experiments to evaluate the performance of our model. In

Manuscript received 24 February 2022; accepted 21 June 2022. Date of publication 12 July 2022; date of current version 3 February 2023. This letter was recommended for publication by Associate Editor H. Zha and Editor C. C. Lerma upon evaluation of the reviewers' comments. This work was supported in part by the National Key Research and Development Program of China under Grant 2018AAA0102504, in part by the National Natural Science Foundation of China (NSFC) under Grant 62003073, and in part by the Sichuan Science and Technology Program under Grants 2021YFG0184, 2020YFSY0012, and 2018GZDZX0037. *(Fengjun Mu and Rui Huang contributed equally to this work.) (Corresponding author: Jing Qiu.)*

Fengjun Mu is with the Center for Robotics, School of Automation Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China, and also with the School of Mechanical and Electrical Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China (e-mail: mufengjun260@gmail.com).

Rui Huang, Kecheng Shi, and Hong Cheng are with the Center for Robotics, School of Automation Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China (e-mail: ruihuang@uestc.edu.cn; uestc_skc@163.com; hcheng@uestc.edu.cn).

Jing Qiu is with the School of Mechanical and Electrical Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China (e-mail: qiujing@uestc.edu.cn).

Xin Li is with the Inception Institute of Artificial Intelligence, the Group 42, Abu Dubai 51133, UAE (e-mail: xinli_uestc@hotmail.com).

Our code is available at https://github.com/mufengjun260/Weak6D.

Digital Object Identifier 10.1109/LRA.2022.3190094

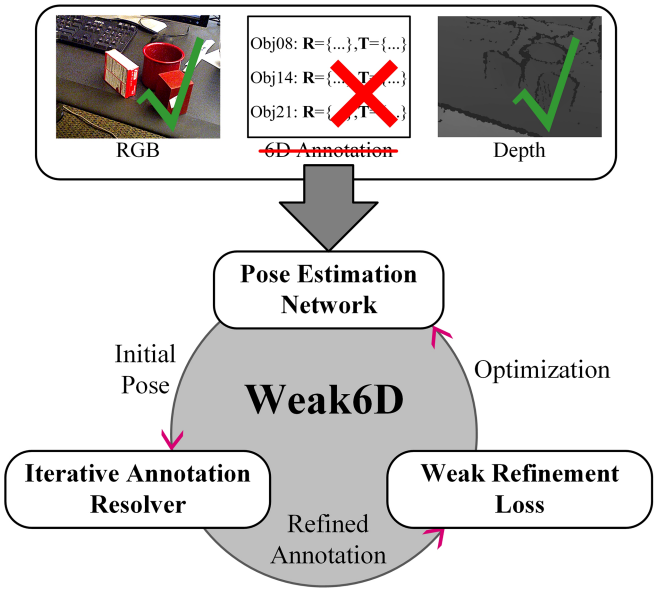

Fig. 1. We develop an end-to-end deep network model '*Weak6D*' for pose estimation without annotated data. Iterative Annotation Resolver refines the initial pose from the network. The refined annotation is employed to Weak Refinement Loss for optimizing the pose estimation network.

the first experiment, the model is trained with only the RGB-D inputs to evaluate performance of the weakly supervised method. Then annotated synthetic data are utilized for initial training, which aiming to prove our model can achieve better performance with annotated synthetic data. Our work is compared with the stat-of-the-art method DenseFusion [12]. The results show that our model achieves a practical accuracy in 6D pose estimation.

In summary, the main contributions of this letter are:

- We propose a novel end-to-end model for 6D object pose estimation named *Weak6D*, which employs learnable balance factors to accelerate the weakly-supervised training process.
- We introduce a weakly-supervised training approach based on Iterative Annotation Resolver. The iterative combination can utilize the optimizing target of the pose estimation network.
- Quantitative and qualitative experiments are evaluated on the YCB-Video dataset [8]. The results demonstrate the effectiveness of our model.

II. Related Work

A. Pose Estimation Methods

6D object pose estimation methods can be roughly classified into template-based methods, correspondence-based and end-to-end methods. **Template-based methods** generate templates from collected object images or 3D models with multiple viewpoints, and then find the best match of the input by using a distance measure [5], [6], [15], [16]. These methods use predefined ways to match the template and the input image. Hinterstoisser et al. [5], [6] generate templates based on the gradient of the RGB image and the normal vector of the object surface. This kind of method requires 6D annotated data to build the template matching library. **Correspondence-based methods** resolve object's pose by extracting correspondence between feature domains and model domain [17]–[20]. Traditional correspondence-based methods extract scale invariant points of interest, such as SIFT [21] or SURF [22]. Recently, learning-based correspondence extractors based on deep neural networks are widely used. Tekin et al. [17] propose a single-shot deep CNN architecture, which can directly detect the 3D bounding box of objects without posterior refinement. Peng et al. [20] regresses pixel-level key point vectors for voting, which could improve the performance in cases of object occlusions. Furthermore, many correspondence-pose resolvers are proposed to estimate the object's 6D poses, such as RANSAC[23], P*n*P [24], and the singe-stage resolver [18]. However, these models are sensitive to the false detection of easily disturbed sparse features, which can not achieve robust pose estimation in complex scenarios. With the development of deep neural network methods, **end-to-end methods** have gradually become popular in object pose estimation [7], [9], [25]–[28]. Yu et al. [8] propose a CNN-based architecture that can directly regress object poses from RGB image inputs. Wang et al. [12] propose DenseFusion, which integrates the pixel-wise dense feature into a global feature to achieve better performance in scenarios with object occlusions. Park et al. [9] learn the 3D representations of the object based on labeled RGB images, which can be extended to unseen objects. Wada et al. [13] further integrate the occupation information into the dense feature, and employ a collision-based optimization method to improve the accuracy of pose estimation in cluttered scenes. These models need amount of annotated datasets to train a deep model with accurate pose estimation. Additionally, some novel methods are trying to eliminate the reliance on annotated data. Wang et al. [29] propose using a differentiable renderer [30] for visually and geometrically based self-supervised processes. Sock et al. [31] employ the wrap-align stage for better establishing constraints between different frames. These self-supervised methods require a differentiable renderer to obtain the optimizing target, which is high computing resource consumption. Zhang et al. proposed *DAKDN* [32], which uses the domain-invariant geometry structure among keypoints to cross the domain gap. Li et al. [33] compute the projected R-IoU loss from multiple views using relative camera poses and 3-D bounding boxes. However, as far as we know, existing object pose estimation methods were verified in slightly occluded LineMod [5] and T-Less [34] dataset, and were not optimized for heavily occluded scenes in YCB-Video, which is more common in robotics applications.

B. Datasets for Pose Estimation

Several datasets are annotated with ground-truth of 6D object poses, for training models with data-driven methods. LineMOD [5] provides manual annotations for around 1,000 images for each of the 15 objects in the dataset. To avoid annotating all frames manually, YCB-Video [8] only annotate object poses

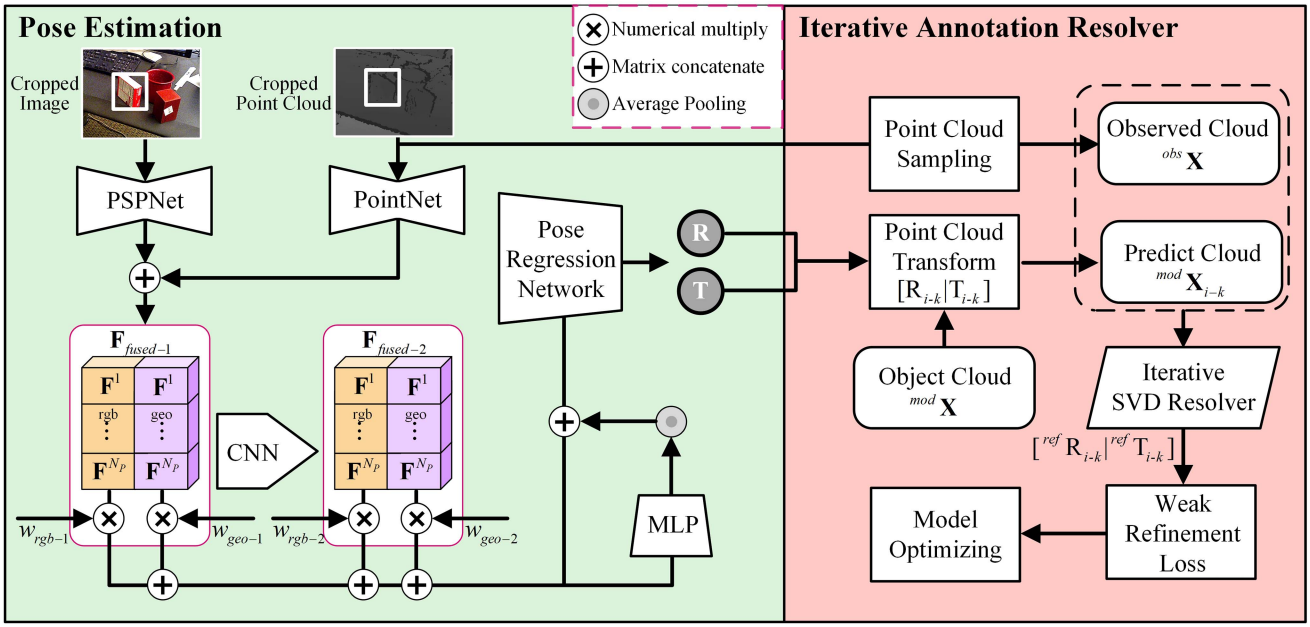

Fig. 2. Overview of the proposed 6D pose estimation model. Our model generates features from RGB-D input independently and fuse them at pixel level. Iterative Annotation Resolver utilizes the refined annotation for pose prediction. We employ the weak refinement loss to optimize the pose estimation network without annotated data.

of the first frame, and uses the camera trajectory to generate each frames' annotations with global optimization. YCB-Occlusion [35] automatically adds heavy occlusions on the basis of YCB-Video for evaluating the robustness of the model. BlenderProc4BOP [36], [37] provides a physically-based renderer (PBR) to generate photorealistic images and the corresponding annotations.

C. Pose Refinement Methods

Pose refinement methods are widely employed to further improve performances of the end-to-end methods. Most methods [8], [27], [28] use Iterative Closest Point (ICP) algorithm as the refinement approach. DenseFusion proposed a neural network-based iterative refinement method that has higher efficiency and performance. MoreFusion [13] introduced collision-based pose refinement to optimize the poses of multiple objects with collision checking jointly. However, existed iterative refinement methods always needs accurate poses from the optimized model, which leads amount of annotated data are necessary to optimize the model.

This letter introduce a novel pose estimation method *Weak6D*, which fuses the refinement into the model training process. Generated refined annotations are used for parameter optimization iteratively. The proposed method can estimate the 6D object poses without manual annotations.

III. Methods

Our goal is to estimate the 6D pose of a set of known objects present in RGB-D images. The 6D pose is presented as a homogeneous transformation matrix $\mathbf{P} = [\mathbf{R} \mid \mathbf{T}] \in SE(3)$, which is composed by a rotation $\mathbf{R} \in SO(3)$ and a translation $\mathbf{T} \in \mathbb{R}^3$. *Weak6D* is designed based on DenseFusion[12], and **Iterative Annotation Resolver** is proposed for dynamically generating the optimization target. In addition, we utilize synthetic data to reduce the requirement of high-cost annotation dependence.

A. Overview of the Model

Fig. 2 illustrates the overview of the proposed *Weak6D*. Our model contains two main stages, the first stage performs semantic segmentation for each object. Then we employ weighted pixel-wise dense fusion to obtain features from each object's cropped image and cropped point cloud. Finally, a pose regression network is employed to regress the object's 6D pose from the feature. Different from existed representative network presented in DenseFusion [12], we employ a weighted feature fusion method for the weak supervision mechanism. In the second stage, an Iterative Annotation Resolver are proposed to generates refined 6D annotation iteratively for weakly supervised training. The details of our model are described below.

B. Semantic Segmentation

A Semantic segmentation network is employed to obtain the interest area of each object. In detail, the RGB image is feeded into an encoder-decoder network to generate semantic segmentation results, then the output map M_{seg} (size $H * W * (N_c + 1)$) for N_c classes is transformed to object's class label, which aiming to generate the binary masks M_{obj-c}(size $H * W$) of the cth object. Since semantic segmentation is not the research point of Weak6D, we directly employ SegNet [38] to obtain the corresponding mask.

C. Pose Estimation

The key technical challenge of pixel-wise fusion module is how to extract more useful information from the RGB-D image. One way is directly fusing the features extracted from cropped color image and the depth image. However, in the industrial scenes, occlusions and segmentation errors would lead to features of low quality, which reduce the accuracy of pose estimation. Therefore, we randomly sample N_p pixels and their corresponding 3D points to generate the pixel-wise dense feature.

As shown in Fig. 2, we firstly clip the color images and depth images to obtain the object's corresponding area. The clipping process is based on the results of semantic segmentation. Then the color feature (size $H * W * D_{color}$) and the geometric feature (size $H * W * D_{geo}$) are extracted by PSPNet [39] and PointNet [40] independently. Here D_{color} and D_{geo} represent output dimensions of color features and geometric features respectively.

Different from DenseFusion [12], we employ learnable balance factors W_b to weight the features before the fusing process. The weighted factors have the ability to balance the RGB features and the geometric features in pose regression network. During the fusion process, w_{rgb-i} and w_{geo-i} are multiplied to the ith fusion layer from the RGB and geometric branches. The fused jth pixel-wised feature can be calculated as follows:

$$\mathbf{F}^{j}_{fused-i} = w_{rgb-i}\mathbf{F}^{j}_{rgb-i} \oplus w_{geo-i}\mathbf{F}^{j}_{geo-i}, \tag{1}$$

where i in $\{1, 2\}$ and $\oplus$ denotes tensor concatenate to operate pixel-wise fusion of different features. F_{rgb-i} and F_{geo-i} indicate the ith RGB and geometric features from different layers respectively. Then, $\mathbf{F}^{j}_{fused} = \mathbf{F}^{j}_{fused-1} \oplus \mathbf{F}^{j}_{fused-2}$ will be utilized for pose regression.

So far we have obtained pixel-wise dense feature from RGB-D inputs. Nevertheless, the pixel-wise dense feature only contains local features. In order to include the global information, we employ a 2-layers shared MLP network and average-pooling to generate a global feature from the pixel-wise dense feature. The pixel-wise and global features are concatenated and feeded into three independent 4-layers CNN. Three independent convolutional networks regress rotation $\mathbf{R}$, translation $\mathbf{T}$, and confidence c for N_p sampled pixels. Finally, we select the sampling point with optimal confidence based on the output of pose regression network, which take its corresponding rotation quaternion and translation as the final estimated 6D poses.

D. Iterative Annotation Resolver

Deep neural networks achieve accurate object pose estimation based on end-to-end training methods, in which iterative refinement methods are widely employed to refine the estimated pose. However, the training process of these networks highly relies on annotated ground truth. Moreover, existing refinement methods always optimized based on accurate initial pose, which is unable to obtain if the models are not finely trained.

To deal with above problems, we present a novel Iterative Annotation Resolver to train the model from captured RGB-D images directly. Compared with the existing ICP-based refinement method, Iterative Annotation Resolver presents a parallel approach based on per-pixel regressed object pose. Iterative Annotation Resolver can dynamically generate the optimizing target of learning progress. Fig. 3 illustrates the architecture of the proposed Iterative Annotation Resolver. Based on the pose estimation network we introduced before, the ith cycle's predicted pose from kth pixel $\mathbf{P}_{i-k} = [\mathbf{R}_{i-k}|\mathbf{T}_{i-k}]$ is firstly obtained from cropped RGB-D input. Since the network at the initial process is not trained and the parameters are initialized using random initialization methods [41], $\mathbf{P}_{0-k}$ can be considered as an inaccurate initial pose. The the cropped depth image is utilized to restore the observed 3D point cloud of the object ${}^{obs}\mathbf{X}$ (size M), by using reverse projection based on intrinsic parameters of the camera.

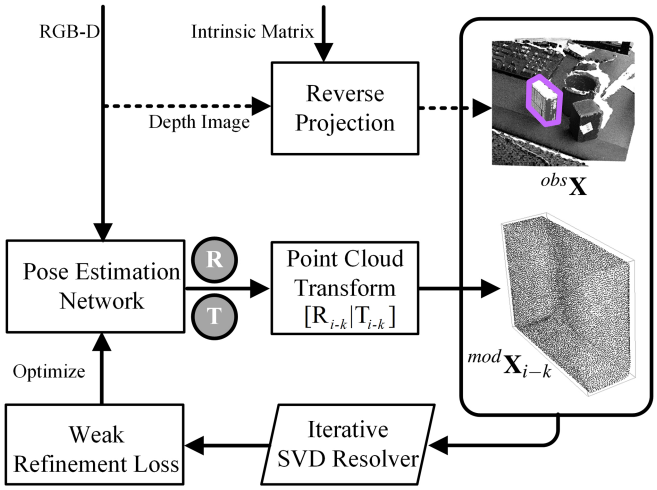

Fig. 3. Architecture of the Iterative Annotation Resolver. We employ the reverse projection for depth image to obtain the object's observed point cloud. Object's sampled point cloud is transformed by network's predicted pose. Finally we utilize the Iterative SVD Resolver to refine the predicted object's pose and optimize the network with Weak Refinement Loss.

The proposed Iterative Annotation Resolver contains two steps. In ith iteration cycle, the first step is to transform the sampled model ${}^{mod}\mathbf{X}$ (size N) with the kth predicted pose as follows:

$${}^{mod}\mathbf{X}_{i-k} = \mathbf{R}_{i-k} \cdot {}^{mod}\mathbf{X} + \mathbf{T}_{i-k}, \tag{2}$$

then use the iterative closest point method to resolve approximate transformation ${}^{ref}\mathbf{P}_{i-k} = [{}^{ref}\mathbf{R}_{i-k}|{}^{ref}\mathbf{T}_{i-k}]$. The refinement process is defined as follows:

$$\min_{{}^{ref}\mathbf{P}_{i-k}} dis = \frac{1}{N}\sum_{j}\min_{h} \|{}^{ref}\mathbf{R}_{i-k} \cdot {}^{mod}\mathbf{X}^{j}_{i-k} + {}^{ref}\mathbf{T}_{i-k} - {}^{obs}\mathbf{X}^{h}_{k}\|^2_2, \tag{3}$$

where $j \in (0, N]$ and $h \in (0, M]$. We use SVD algorithm [42] iteratively for several times (ω_1) to optimize dis.

The second step of the ith iteration cycle is to optimize model parameters using **Weak Refinement Loss**. Traditional loss functions are defined as the distance between the object's sampled model in ground truth 6D annotated pose and corresponding

points transformed by the predicted pose. In this letter, we compute the loss function without 6D annotated pose and rendered images. As shown in (3), we use the distance between ${}^{mod}\mathbf{X}$ and refined point cloud. Specifically, the loss which minimizes for the prediction kth pixel is defined as follows:

$$L_{i-k}^{p} = \frac{1}{N} \sum_{j \in (0,N]} \| {}^{ref}\mathbf{R}_{i-k} \cdot {}^{mod}\mathbf{X}_{i-k}^{j} + {}^{ref}\mathbf{T}_{i-k} - {}^{mod}\mathbf{X}_{i-k}^{j} \|_2. \quad (4)$$

In the loss function of existed pose estimation methods [8], symmetrical objects are specifically considered to avoid the impression of rotational symmetry. In the process of iteratively solving ${}^{ref}\mathbf{P}_{i-k}$, the nearest neighbor method has been utilized to minimize dis, which leads L_{i-k}^{p} is robust to the symmetric objects.

In order to balance the confidence among the per dense-pixel predictions, we weighted the distance with pixel-wise confidence c_{i-k} with added a confidence regularization term:

$$L = \frac{1}{N} \sum_{j \in (0,N]} \left(L_{i-k}^{p} \cdot c_{i-k}^{j} - \omega_2 \cdot \log\left(c_{i-k}^{j}\right) \right). \quad (5)$$

With the proposed Iterative Annotation Resolver with weak refinement loss, the model parameters can be optimized iteratively through above two steps for accurate object pose estimation. Even though this iterative optimization method is relatively simple, but *Weak6D* employed a pixel-wise feature-based pose estimation network. The optimization process is based on massive sampled pixels from large-scale input RGB-D images, and can implicitly establish connections among single discrete optimization.

IV. Experiments

In the experimental section, we evaluate our model on the YCB-Video dataset [8]. Since there are few SOTA works based on unannotated data, we choose to compare with the supervised learning method DenseFusion. First of all, we train *Weak6D* with only the RGB-D videos to verify the performance of our method without 6D annotations. Then, we use annotated synthetic data for initial training with the loss from ground truth [12], and further fine-tune *Weak6D* using the Iterative Annotation Resolver, which aiming to evaluate the performance of proposed *Weak6D* with mixture of annotated synthetic data and unannotated RGB-D data.

A. Datasets

YCB-Video dataset [8] is constructed based on 21 YCB objects [43], which contains 130 thousands frames of RGB-D data collected from 92 real captured videos and 80 thousands frames of synthetic RGB-D data. YCB-Video dataset has been commonly used for evaluation of 6D pose estimation in prior work, since cluttered scenes increase the difficulty of accurate pose estimation. LineMod dataset [5] offers a limited data scale for weakly-supervised learning. Therefore, we only experiment on YCB-Video dataset [8] with large-scale RGB-D data.

In *Weak6D*'s experiment that completely uses the weakly supervised method for training, we only retained the RGB-D data of the real recorded part of the YCB-Video dataset. In the experiment combining synthetic data for training, we additionally used synthetic data with annotation information.

B. Metrics

1) Metrics for Accuracy: We use two accuracy metrics to report on the YCB-Video dataset. The first one is the area under the ADD-S curve (AUC) following PoseCNN [8], where the maximum threshold of AUC is set to $0.1m$.

$$ADD - S = \frac{1}{N} \sum_{x_1 \in {}^{mod}\mathbf{X}} \min_{x_2 \in {}^{mod}\mathbf{X}} \| (\mathbf{R}x_1 + \mathbf{T}) - \left(\tilde{\mathbf{R}}x_2 + \tilde{\mathbf{T}}\right) \|, \quad (6)$$

As described in (6), given the estimated pose $\tilde{\mathbf{P}}$ and ground truth pose $\mathbf{P}$, ADD-S calculates the mean distance from each 3D model point transformed by $\tilde{\mathbf{P}}$ to its closest neighbour on the target model transformed by $\mathbf{P}$. The second metric for accuracy is the percentage of ADD-S smaller than 2 cm (< 2 cm), which measures the estimations under the minimum tolerance for robot grasping and manipulation.

C. Implementation Details

Semantic segmentation results used in the experiments are generated by a SegNet semantic segmentation model [38]. We set $\omega_1 = 20$ to perform multiple iterative optimization using SVD for better refined weakly annotation, and a high-performance GPU parallelization implementation is developed to avoid slowing down the training speed. We follow the hyper-parameters setting $\omega_2 = 0.015$ in DenseFusion. The Adam optimizer is employed to optimize the models in the training processes. ICP stage in *PureICP* test and refinement process is set to 50 iterations.

D. Experimental Results

Table I shows the evaluation results of *Weak6D* and other methods for all 21 objects in YCB-Video dataset. *PureICP* To ensure a fair comparison, all methods use the same segmentation masks. Compared with DenseFusion, our model achieves the same excellent results as DenseFusion with weakly supervised learning.

1) Effect of Weakly Supervised Learning: We use the RGB-D data from YCB-Video's training set for weakly optimizing *Weak6D*. After the training process with Iterative Annotation Resolver and Weak Refinement Loss, our method obtains 98.4% and 98.5% under different hyper-parameters on YCB-Video dataset. With an additional ICP refinement process, *Weak6D* obtains 80.4% on AUC matrics, which is 9.1% higher than directly using the ICP method to estimate the poses (*PureICP*), which proves that *Weak6D* indeed enables the network to learn the knowledge related to pose estimation from unannotated data. We analyzed this phenomenon for some assumptions. The weakly-supervised training with Iterative Annotation Resolver

TABLE I
QUANTITATIVE EVALUATION OF 6D POSE ON YCB-VIDEO DATASET

	PureICP		DenseFusion		Weak6D					
					ω_1=1		ω_1=20		ω_1=20, with ICP	
Object	AUC	$< 2cm$	AUC	$< 2cm$	AUC	$< 2cm$	AUC	$< 2cm$	AUC	$< 2cm$
002_master_chef_can	100.0	74.3	99.9	98.8	100.0	77.0	100.0	89.2	100.0	97.3
003_cracker_box	100.0	45.1	100.0	99.3	100.0	10.4	100.0	0.6	100.0	41.8
004_sugar_box	100.0	82.3	99.9	98.6	100.0	64.0	99.8	69.7	100.0	84.6
005_tomato_soup_can	100.0	88.8	99.9	99.1	99.9	98.2	99.4	92.5	100.0	90.1
006_mustard_bottle	100.0	66.3	100.0	100.0	100.0	100.0	100.0	100.0	100.0	94.4
007_tuna_fish_can	100.0	87.2	100.0	98.4	100.0	100.0	99.9	99.7	100.0	92.8
008_pudding_box	100.0	100.0	100.0	100.0	100.0	95.3	100.0	99.1	100.0	100.0
009_gelatin_box	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
010_potted_meat_can	94.5	75.9	99.2	95.8	96.2	91.6	99.5	82.5	98.7	73.9
011_banana	100.0	100.0	100.0	97.1	100.0	41.7	100.0	58.8	100.0	100.0
019_pitcher_base	100.0	34.9	100.0	100.0	100.0	89.3	100.0	78.4	100.0	44.8
021_bleach_cleanser	99.4	56.0	99.9	98.9	100.0	86.9	100.0	75.1	98.6	76.6
024_bowl	99.2	49.2	100.0	72.2	100.0	1.7	100.0	0.7	100.0	52.2
025_mug	98.1	63.8	100.0	98.6	99.8	100.0	100.0	99.8	96.7	69.0
035_power_drill	100.0	52.3	100.0	99.5	100.0	36.1	100.0	46.1	100.0	77.9
036_wood_block	100.0	15.0	100.0	84.9	100.0	48.3	100.0	23.5	100.0	2.7
037_scissors	100.0	97.6	100.0	98.9	100.0	74.0	99.4	80.6	100.0	100.0
040_large_marker	100.0	100.0	100.0	99.8	100.0	80.7	99.5	88.7	100.0	100.0
051_large_clamp	100.0	90.0	73.7	59.0	71.7	22.5	73.1	67.4	100.0	90.5
052_extra_large_clamp	100.0	73.0	99.7	11.2	99.3	3.5	99.5	1.5	100.0	81.6
061_foam_brick	100.0	96.0	100.0	99.7	99.7	99.7	94.8	98.9	100.0	93.7
MEAN	99.5	71.3	98.7	92.5	98.4	68.7	98.5	69.7	99.7	80.4

TABLE II
THE EFFECT OF TRAINING WITH AND WITHOUT SYNTHETIC DATA

Weak6D			
	ADD	ADD-S	$< 2cm$
only *Weak6D*	71.7	98.5	69.7
only synthetic data	83.6	84.6	70.2
synthetic data & *PureICP*	94.0	99.4	82.3
synthetic data & *Weak6D*	97.6	99.6	84.6

TABLE III
ABLATION STUDY ON YCB-VIDEO

	ADD	ADD-S	$< 2cm$	epochs
original *Weak6D*	73.2	98.1	63.6	22
w/o learnable factors	69.6	98.5	47.5	38

will simultaneously optimize the result from sampled pixels, and the traversal on dataset-scale unannotated data will implicitly establish connections among single discrete optimization. In an abstract, the weakly-supervised learning progress jointly refines the unannotated RGB-D images at the dataset level, which can achieve significantly better results than single refinement.

However, since 6D annotated data is not utilized, the refined annotation is generated based on initial pose and object's observed point cloud. With the accuracy improves of the pose estimation network, quality of refined annotations will be negatively affected by factors such as distance measurement error of the depth camera, the point cloud's atrous, etc., resulting that in AUC(< 2 cm), *Weak6D* got worse results than DenseFusion. Even though the limitation of the ICP algorithm makes it difficult for higher accuracy, *Weak6D* still has the practical significance in industrial scenarios.

2) Hybrid Training Performance: We firstly use annotated synthetic data to initialize the network's parameters, and performs weakly supervised training with the proposed Iterative Annotation Resolver. Table II compares the performance of *Weak6D* trained with unannotated RGB-D data and with extra synthetic data. By employing synthetic data, ADD and ADD-S(< 2 cm) increased by 25.9% and 1.1%(14.9%) respectively with refinement process. Sim2Real often suffers from poor performance when training with synthetic data, and the sensor noise in the real collected unannotated data will reduce the accuracy of *Weak6D*'s refined annotation. Therefore, we combined using both refined annotation and annotated synthetic data to train *Weak6D* for performance.

3) Ablation Study: We present ablation experiments on YCB-Video dataset. We removed learnable balance factor from *Weak6D* and retrained the models under same config. Table III shows the comparison result between original *Weak6D* and modified model. The mechanism only employ 4 addition parameters, and brings apparent improvement on the accuracy matrix, especially on ADD-S(< 2 cm), and the convergence speed increased by 42.1%. *Weak6D* takes only 33 hours for training with single NVIDIA 2080Ti.

4) Iterative Annotation Resolver: In Table I, we conducted weakly supervised training of the network under different settings of omega $\{1, 20\}$. The training results show that the difference in ω_1 does not have a big impact on the optimization speed of the network over time. Our analysis of this phenomenon is that a smaller omega will cause the two-step method of "generating weak annotation" and "fine tune the network to fit weak annotation" to be iterated more frequently, while a larger omega will make the generated refined annotation more accuracy, so the network optimization process will be more accurate.

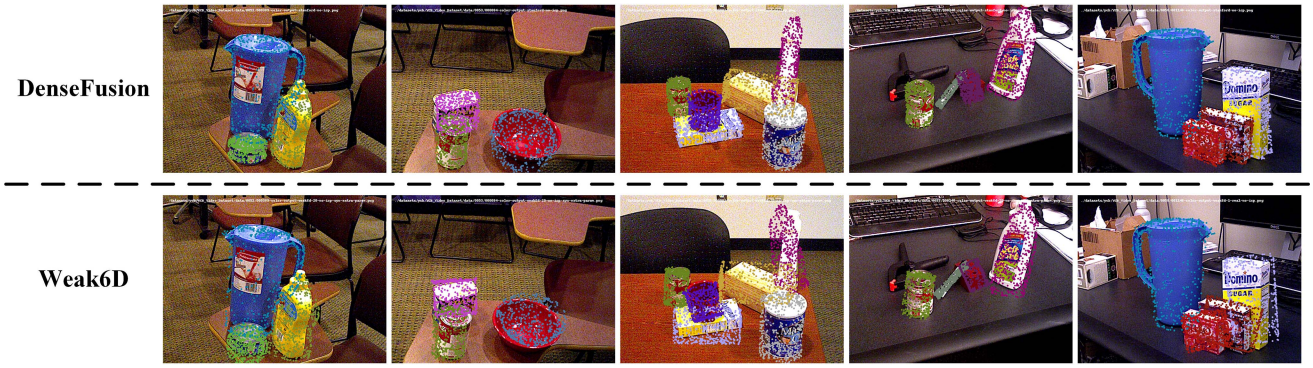

Fig. 4. Qualitative results on the YCB-Video dataset. Two methods shown here are tested with the same segmentation masks. Each object point cloud is transformed with the estimated pose and then projected to the 2D image. Although the training proceed without annotated data, the performance of our model is close to the state-of-the-art method DenseFusion.

5) Qualitative Evaluation: Fig. 4 visualizes some samples estimations made by DenseFusion and *Weak6D*. As we see in the upper line of Fig. 4, *Weak6D* trained with unannotated data reached the effect of supervising trained DenseFusion. As depicted in the bottom line, the estimating results of *Weak6D* have some minor errors in scenarios with heavy occlusions but still within an acceptable range. Thanks to the Iterative Annotation Resolver and Weak Refinement Loss, our model achieves accurate pose estimation even without annotated data.

V. Conclusion

This letter presents an end-to-end model named *Weak6D*, which could be learned with unannotated RGB-D data. The core of the proposed approach is the novel optimizing method Iterative Annotation Resolver, which has the ability to utilize the captured RGB-D data for model training directly. Furthermore, we employ a weak refinement loss to optimize the pose estimation network with refined object poses. We evaluated the proposed *Weak6D* in the YCB-Video dataset, and experimental results show our model achieved practical results without annotated data.

References

[1] Y. Yang, H. Liang, and C. Choi, "A deep learning approach to grasping the invisible," *IEEE Robot. Automat. Lett.*, vol. 5, no. 2, pp. 2232–2239, Apr. 2020.

[2] K. Fang et al., "Learning task-oriented grasping for tool manipulation from simulated self-supervision," *Int. J. Robot. Res.*, vol. 39, no. 2/3, pp. 202–216, 2020.

[3] B. Fang et al., "A dual-modal vision-based tactile sensor for robotic hand grasping," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2018, pp. 4740–4745.

[4] W. Fang et al., "Visual-guided robotic object grasping using dual neural network controllers," *IEEE Trans. Ind. Inform.*, vol. 17, no. 3, pp. 2282–2291, Mar. 2021.

[5] S. Hinterstoisser et al., "Multimodal templates for real-time detection of texture-less objects in heavily cluttered scences," in *Proc. Int. Conf. Comput. Vis.*, 2011, pp. 858–865.

[6] S. Hinterstoisser et al., "Model based training, detection and pose estimation of texture-less 3D objects in heavily cluttered scenes," in *Proc. 11th Asian Conf. Comput. Vis. - Vol. Part I*, 2012, pp. 548–562.

[7] E. Brachmann, A. Krull, F. Michel, S. Gumhold, J. Shotton, and C. Rother, "Learning 6D object pose estimation using 3D object coordinates," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 536–551. .

[8] Y. Xiang, T. Schnidt, V. Narayanan, and D. Fox, "PoseCNN: A convolutional neural network for 6D object estimation in cluttered scences," in *Proc. Robot.: Sci. Syst.*, 2018. [Online]. Available: http://www.roboticsproceedings.org/rss14/p19.html

[9] K. Park, A. Mousavian, Y. Xiang, and D. Fox, "LatentFusion: End-to-end differentiable reconstruction and rendering for unseen object pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 10707–10716.

[10] C. Li, J. Bai, and G. D. Hager, "A unified framework for multi-view multi-class object pose estimation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 263–281.

[11] D. Xu, D. Anguelov, and A. Jain, "PointFusion: Deep sensor fusion for 3D bounding box estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 244–253.

[12] C. Wang et al., "DenseFusion: 6D object pose estimation by iterative dense fusion," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 3343–3352.

[13] K. Wada, E. Sucar, S. James, D. Lenton, and A. J. Davison, "MoreFusion: Multi-object reasoning for 6D pose estimation from volumetric fusion," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 14528–14537.

[14] Y. He, H. Huang, H. Fan, Q. Chen, and J. Sun, "FFB6D: A full flow bidirectional fusion network for 6D pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 3002–3012.

[15] Z. Li, G. Wang, and X. Ji, "CDPN: Coordinates-based disentangled pose network for real-time RGB-based 6-DoF object pose estimation," in *Proc. Int. Conf. Comput. Vis.*, 2019, pp. 7677–7686.

[16] Y. Li, G. Wang, X. Ji, Y. Xiang, and D. Fox, "DeepIM: Deep iterative matching for 6D pose estimation," in *Proc. 15th Eur. Conf. Comput. Vis.*, 2018, pp. 695–711.

[17] B. Tekin, S. N. Sinha, and P. Fua, "Real-time seamless single shot 6D object pose prediction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 292–301.

[18] Y. Hu, P. Fua, W. Wang, and M. Salzmann, "Single-stage 6D object pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 2927–2936.

[19] C. Song, J. Song, and Q. Huang, "HybridPose: 6D object pose estimation under hybrid representations," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 428–437.

[20] S. Peng, Y. Liu, Q. Huang, X. Zhou, and H. Bao, "PVNet: Pixel-wise voting network for 6DoF pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 4561–4570.

[21] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.

[22] H. Bay, T. Tuytelaars, and L. V. Gool, "SURF: Speeded up robust features," in *Proc. Eur. Conf. Comput. Vis.*, 2006, pp. 404–417.

[23] M. A. Fischler and R. C. Bolles, "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," *Commun. ACM*, vol. 24, no. 6, pp. 381–395, 1981.

[24] V. Lepetit, F. Moreno-Noguer, and P. Fua, "EPnP: An accurate O(n) solution to the PnP problem," *Int. J. Comput. Vis.*, vol. 81, no. 2, pp. 155–166, 2009.

[25] B. Drost, M. Ulrich, N. Navab, and S. Ilic, "Model globally, match locally: Efficient and robust 3D object recognition," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2010, pp. 998–1005.

[26] J. Shotton, B. Glocker, C. Zach, S. Izadi, A. Criminisi, and A. W. Fitzgibbon, "Scene coordinate regression forests for camera relocalization in RGB-D images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 2930–2937.

[27] M. Sundermeyer, Z. Marton, M. Durner, M. Brucker, and R. Triebel, "Implicit 3D orientation learning for 6D object detection from RGB images," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 712–729.

[28] 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 *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 1530–1538.

[29] G. Wang, F. Manhardt, J. Shao, X. Ji, N. Navab, and F. Tombari, "Self6D: Self-supervised monocular 6D object pose estimation," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 108–125.

[30] W. Chen et al., "Learning to predict 3D objects with an interpolation-based differentiable renderer," in *Proc. 33rd Int. Conf. Neural Inf. Process. Syst.*, 2019, pp. 9609–9619.

[31] J. Sock, G. Garcia-Hernando, A. Armagan, and T.-K. Kim, "Introducing pose consistency and warp-alignment for self-supervised 6D object pose estimation in color images," in *Proc. Int. Conf. 3D Vis.*, 2020, pp. 291–300.

[32] S. Zhang, W. Zhao, Z. Guan, X. Peng, and J. Peng, "Keypoint-graph-driven learning framework for object pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 1065–1073.

[33] F. Li, I. Shugurov, B. Busam, S. Yang, and S. Ilic, "WS-OPE: Weakly supervised 6-D object pose regression using relative multi-camera pose constraints," *IEEE Robot. Automat. Lett.*, vol. 7, no. 2, pp. 3703–3710, Apr. 2022.

[34] T. Hodan, P. Haluza, Š. Obdržálek, J. Matas, M. Lourakis, and X. Zabulis, "T-LESS: An RGB-D dataset for 6D pose estimation of texture-less objects," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2017, pp. 880–888.

[35] F. Mu, R. Huang, A. Luo, X. Li, J. Qiu, and H. Cheng, "TemporalFusion: Temporal motion reasoning with multi-frame fusion for 6D object pose estimation," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2021, pp. 5930–5936.

[36] M. Denninger, M. Sundermeyer, D. Winkelbauer, D. Olefir, and T. Hodan, "BlenderProc: Reducing the reality gap with photorealistic rendering," in *Proc. Robot.: Sci. Syst. (RSS) Workshops*, 2020. [Online]. Available: https://sim2real.github.io/assets/papers/2020/denninger.pdf

[37] M. Denninger et al., "BlenderProc," 2019, *arXiv:1911.01911*.

[38] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017.

[39] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 6230–6239.

[40] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "PointNet: Deep learning on point sets for 3D classification and segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 77–85.

[41] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1026–1034.

[42] K. S. Arun, T. S. Huang, and S. D. Blostein, "Least-squares fitting of two 3-D point sets," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-9, no. 5, pp. 698–700, Sep. 1987.

[43] 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 *Proc. Int. Conf. Adv. Robot.*, 2015, pp. 510–517.