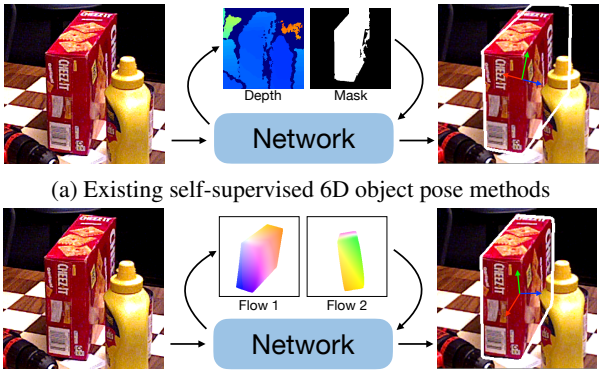
Pseudo Flow Consistency for Self-Supervised 6D Object Pose Estimation

Anonymous ICCV submission

Paper ID 8656

Abstract

Recent self-supervised 6D object pose methods have shown fast progress. However, most can only work with additional depth information or rely on the accurate annotation of 2D segmentation masks, limiting their application range. In this paper, we propose a 6D object pose method that can be trained with pure RGB images having neither 2D nor 3D annotations. We first propose obtaining a rough pose initialization by networks trained on synthetic images rendered from the target's 3D mesh. Then, we introduce a refinement strategy leveraging the geometry consistency in synthetic-to-real image pairs after some simple data augmentations. The core of our method is the exploitation of the geometry constraint across unannotated images, which we propose to formulate as pixel-level flow consistency between the training images with dynamically generated pseudo labels. We evaluate our method on three challenging datasets and demonstrate that it outperforms most state-of-the-art self-supervised methods significantly, with little auxiliary information.



(a) Existing self-supervised 6D object pose methods

(b) Our solution based on pseudo flow consistency

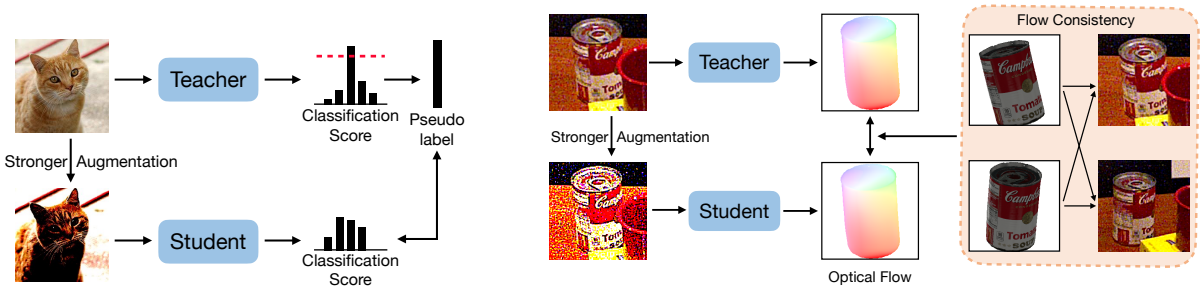
Figure 1. **Comparison of self-supervised object pose methods. (a)** Most existing self-supervised object pose methods rely on either the depth image [51, 50, 4, 28] or additional mask annotations [57, 42], limiting their application range. **(b)** In contrast, our method can be trained only with the guidance of flow consistency from the intrinsic geometry constraint across multiple different views, and produces more accurate results than existing solutions, without relying on any other auxiliary information.

1. Introduction

The problem of 6D object pose estimation is to accurate estimation of the 3D rotation and 3D translation of a rigid object with respect to the camera. This 6D pose estimates give essential information about the world beyond classical 2D understanding and became a fundamental component in many applications, such as robotic manipulation [5], autonomous driving [33], and augmented reality [34].

Recent progress in this field has significantly improved the robustness and accuracy of the model [45, 56, 18, 52, 7, 26, 16, 15]. Most of these approaches rely on a large number of real images with accurate 6D pose annotations. But, compared to classical 2D annotation however, these 6D annotations are either very hard to obtain [35, 31] or are prone to large labeling errors [13, 9, 54]. Most recent methods propose to use techniques based on image synthesis [12] or self-supervised learning [51, 50] to handle this problem. The main problem with synthetic images is the large domain gap to the real images, making the generalization suboptimal [39, 56], while most self-supervised methods rely on additional information. Some can only work with additional depth images [51, 50, 28, 4] or others needs pixel-level annotation in a segmentation masks [42, 57], which prevents the general applicability, as shown in Fig. 1.

In this work, we propose a self-supervised framework for 6D object pose estimation, which relies on neither depth nor additional 2D annotations. We first generate a synthetic dataset based on rendered images from the target's 3D mesh and train networks only on this dataset to get an initial pose initialization. To close the domain gap between the synthetic and real data, we use a refinement strategy where we compare the rendered reference image according to the initial pose and the real input based on pseudo labels [43, 58]. Although widely used in other fields [48, 55, 10, 8], the two fundamental problems of pseudo labeling, including the generative strategy of creating pseudo labels and the selection strategy of choosing high-quality labels from all noisy



(a) The standard method for classification (b) Our method for object pose estimation

Figure 2. **Different self-supervised methods. (a)** The teacher-student learning scheme is a classical framework for self-supervised classification [48]. The key is how to determine the criteria that can select high-quality pseudo label candidates from the noisy predictions of the teacher network.For image classification, one can obtain the prediction quality by the output distribution after the softmax operation easily, which is usually implemented by checking if the probability of any class is above a threshold [43, 58]. **(b)** However, there is no such easy way to determine the quality of an object pose prediction without the ground truth. We propose to formulate pseudo object pose labels as pixel-level optical flow supervision signals, and use the flow consistency based on the underlying geometry constraint from multiple views.

generated candidates, are still open questions in 6D object pose estimation.

We propose to formulate the pseudo 6D pose labels as 2D flow supervision signals in a render-and-compare framework [13, 23, 26, 56, 18, 29]. Unlike the common render-and-compare frameworks that need accurate pose annotations, we propose to render multiple images near the initial pose, and compare them with the real input. From that we train the model only with the guidance of flow consistency between these image pairs, making it independent of the actual annotations. Additionally, we use a geometry-based strategy that only selects the most consistent pixels across multiple rendered views to facilitate the network training.

We evaluate our method on three challenging datasets LINEMOD [11], Occluded-LINEMOD [21], and YCB-V [54], and show that it outperforms state-of-the-art self-supervised methods significantly, including those methods relied on depth image [51, 50] or auxiliary annotation information [57, 42].

Our contributions can be summarized as the following. First, we investigate the problem of the standard teacher-student methods in selecting high-quality pseudo labels for self-supervised object pose estimation. Second, we propose a strategy based on flow consistency that embeds the geometry constraint from multiple views. Finally, we demonstrate its effectiveness by significantly outperforming state-of-the-art self-supervised object pose methods, without relying on other auxiliary information.

2. Related Work

Object pose estimation has shown significant progress recently, based on different techniques, such as direct pose regression [52, 7, 1, 23], 2D reprojection regression [40, 39, 15, 16, 37], 3D keypoint prediction [27, 38, 45, 9], and differentiable PnP solver [24, 14, 2, 3]. However, most of these methods rely on a large number of real images with accurate 6D pose annotation, which is usually hard to obtain in practice, especially in cluttered scenes with multiple object instances and occlusions [54, 13].

Some recent methods tackle this problem by training on synthetic images rendered from the target's 3D mesh [9, 39, 29], but this strategy suffers from the domain gap between the synthetic and real image sets [56, 6]. In contrast, other object pose refinement methods have shown significant improvement in the generalization ability across different domains [26, 29, 23, 18, 56] and especially [13], which show comparable results as the state-of-the-art methods with only about one tenth of real images involved in training. Although this promising progress, these pose refinement methods still need many annotated real images for training, and can not easily benefit from more real data without further annotations.

To solve this problem, some recent self-supervised methods [51, 57, 42, 50] try to completely remove the dependency on pose annotations. Most of them are based on a strategy that compares the synthetic image rendered from the intermediate pose with the real image, and backpropagate the gradient through a differentiable renderer to update the network's weights during training, expecting to align the rendered image with the real input without explicit annotations ([20, 30]). This type of strategy, however, relies heavily on the performance of comparing the final rendered image and the real input, which suffers from the domain gap or occlusion in clutter scenes, making them often rely either on depth [51] or on additional pixel-level annotations in segmentation masks [57]. In contrast, we propose to compare multiple real inputs directly with the help of intermediate synthetic images and force networks to learn pixel-level flow consistency between different views to reduce the domain gap problem. Additionally, we remove most of the noisy flow candidates by the intrinsic geometry constraint

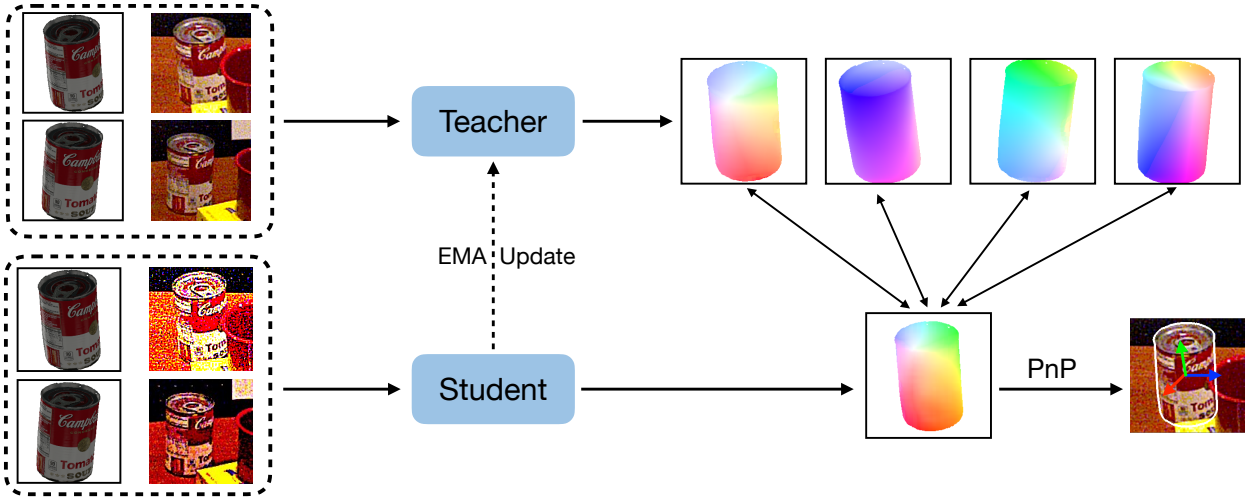


Figure 3. **Method overview.** We first obtain the pose initialization based on a pose network train on synthetic data, and then train our refinement framework on unannotated real training images in a self-supervised manner. Our proposed self-supervised framework is based on a teacher-student learning scheme, where the teacher model is updated using the EMA strategy by the student's weights. Given a rough pose initialization, we render multiple synthetic images around this initial pose, and create multiple image pairs between the synthetic and real images. We propose to use the flow consistency between these image pairs to add geometry constraints to the network training, making it independent of pose annotations.

during training, making it robust to occlusions and independent of manual segmentation masks.

Pseudo labeling is one of the basic techniques used in some recent self-supervised pose methods like [28, 4]. However, these approaches also rely on the additional depth image to select valid pseudo labels, and only update the pseudo labels after finishing the previous training, which usually means the model needs to be trained multiple times to embrace the progressively updated pseudo labels. In contrast, our model only needs to be trained once and does not rely on the auxiliary depth image.

Our method is related to the recent teacher-student formulation of pseudo labeling [43, 58, 48, 55, 22, 59, 25], which works under the assumption that the generated high-quality pseudo label of the teacher can be used to supervise the student network when having the same input as the teacher but only stronger data augmentations. Although this simple general framework has been widely used in image classification [43, 58], object detection [55, 25, 59, 47], and semantic segmentation [22], we find it is sensitive to the quality of generated pseudo labels, and there is not an easy way to select the high-quality pseudo pose labels in the context of 6D object pose estimation. To solve this problem, we propose to formulate pseudo 6D pose labels as pixel-level flow supervision signals and select high-quality pseudo flow labels based on the flow consistency across multiple different views during training. Our experiments show that our method significantly outperforms the baseline teacher-student methods.

3. Approach

Given a dataset of calibrated RGB images and the 3D mesh of the target, our goal is to train a self-supervised model on this dataset to estimate the 6D object pose of the target, without relying on a depth image or any auxiliary information, such as 6D pose and 2D mask annotations. We first create a synthetic dataset by rendering the 3D mesh of the target in different poses and train some existing pose networks [16, 23] on it to obtain a rough pose initialization [13, 26, 18]. The core component of our method is a self-supervised pose refinement framework, which we will discuss in detail in this section. We first show an overview of our self-supervised framework, and then present how we formulate the flow consistency between different views based on the intrinsic geometry constraint. Finally, we show how we extend it to multiple image pairs to further increase the robustness.

3.1. Framework Overview

We use a teacher-student architecture [48] for our self-supervised framework. It contains two networks with identical network structures, but not shared weights, which are called the teacher and student, respectively. During training, when an image input of the teacher network can produce a prediction that can fulfill some criteria, we convert this prediction to a one-hot pseudo-label, and use it to supervise the student network with the same image input but only stronger data augmentations. After the weights updating of the student network by gradient backpropagation super-

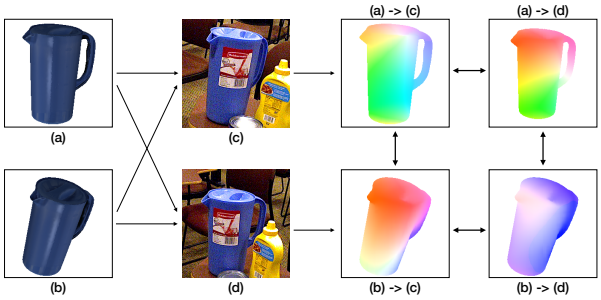


Figure 4. **Illustration of flow consistency.** Using pose initialization P for the target image **(c)**, we render the object under P as **(a)**, and additionally render it under another pose around P, resulting in **(b)**, as well as randomly sampling another real view **(d)**. There are two types of flow consistency. Firstly, the flow predicted from the same synthetic one to the multiple real ones should be consistent, *i.e.*, the flow $(a) \rightarrow (c)$ and the flow $(a) \rightarrow (d)$ should be consistent. Secondly, the flow predicted from the multiple synthetic ones to the same real one should be consistent, *i.e.*, the flow $(a) \rightarrow (c)$ and the flow $(b) \rightarrow (c)$ should be consistent.

vised by this valid pseudo label, we then update the weights of the teacher network by a simple exponential moving averaging (EMA) strategy from the student network:

$$\mathbf{W}_t = \alpha \mathbf{W}_t + (1-\alpha)\mathbf{W}_s, \tag{1}$$

where $\mathbf{W}_t$ and $\mathbf{W}_s$ are the network weight parameters of the teacher and student network, respectively, and α is the exponential factor, which is typically 0.999. The weight updating and pseudo label generation are conducted after each iteration during training, making it much more efficient than other pseudo-label-based object pose methods, which can only produce pseudo labels after the whole training pipeline and need to train the model multiple times [4, 28].

Our main problem is how to determine the criteria that can select high-quality pseudo label candidates from the noisy predictions of the teacher network. For the image classification task as in [48], one can obtain the prediction quality by the output distribution after the softmax operation easily, which is usually implemented by checking if the probability of any class is above a threshold [43, 58]. However, there is no such easy way to determine the quality of an object pose prediction without the ground truth. We discuss our solution in the following sections.

3.2. Flow Consistency across Multiple Views

To solve the problem of difficulties in determining the quality of object pose predictions, we first formulate object pose refinement as a problem of estimating dense 2D-to-2D correspondence, or optical flow estimation, as in PFA [13] which is a fully-supervised object pose method. To tackle the problem of no pose annotation for the computation of ground truth flow, we render multiple images around the initial pose, and predict the flow between each of them and the real input. In principle, since both the rendered images and real input image are 2D reprojections of the same 3D object, the flow prediction that aligns more with the underlying geometry should have a higher probability of being of high quality.

More formally, given an unannotated real image $\mathbf{I}^t$ and the obtained initial pose $\mathbf{P}_0$ from networks trained only on synthetic data, we randomly generate another $n-1$ poses $\{\mathbf{P}_1, \cdots, \mathbf{P}_{n-1}\}$, around the initial pose $\mathbf{P}_0$, and then create n synthetic images by rendering the target under the corresponding poses, generating n image pairs:

$$\{(\mathbf{I}_i^r, \mathbf{I}^t)\}, \quad 0 \leq i \leq n-1, \tag{2}$$

where $\mathbf{I}_i^r$ is the rendered image of the target under pose $\mathbf{P}_i$.

For an object having N 3D keypoints, the 2D reprojection of a 3D keypoint $\mathbf{p}_j, 1 \leq j \leq N$, under pose $\mathbf{P}_i$ can be obtained by

$$\lambda_{ij}^r \begin{bmatrix} \mathbf{u}_{ij}^r \\ 1 \end{bmatrix} = \mathbf{K}(\mathbf{R}_i \mathbf{p}_j + \mathbf{t}_i), \tag{3}$$

where λ_{ij}^r is a scale factor, $\mathbf{u}_{ij}^r$ is the 2D image location, $\mathbf{K}$ is the intrinsic camera matrix, and $\mathbf{R}_i$ and $\mathbf{t}_i$ are the rotation and translation of pose $\mathbf{P}_i$, respectively. We then establish 3D-to-2D correspondence $\mathbf{p}_j \leftrightarrow \mathbf{u}_{ij}^r$ under pose $\mathbf{P}_i$. For the real image $\mathbf{I}^t$, although its true pose $\mathbf{P}^t$ is unknown to us, the relation between the 3D keypoint $\mathbf{p}_j$ and its 2D image location $\mathbf{u}_j^t$ still follows the perspective principle of Eq. 3, implicitly generating the correspondence $\mathbf{p}_j \leftrightarrow \mathbf{u}_j^t$.

We train a network to predict dense 2D-to-2D correspondence $\mathbf{F}_i^{r \rightarrow t}$ between the two images in each image pair of Eq. 7, such that

$$\mathbf{u}_{ij}^r + \mathbf{f}_i^{r \rightarrow t} = \mathbf{u}_j^t \ , \tag{4}$$

where $\mathbf{f}_i^{r \rightarrow t}$ is the corresponding 2D flow vector. Although $\mathbf{u}_j^t$ is unknown during training, we still obtain the geometry constraint that the 2D image locations $\{\mathbf{u}_{ij}^r + \mathbf{f}_i^{r \rightarrow t}\}$ of the same 3D keypoint $\mathbf{p}_j$ from different synthetic views $1 \leq i \leq n$ should be the same.

We use the standard variance of all the predicted $\mathbf{u}_j^t$ from different views to determine if the current pixel's flow prediction is a valid pseudo label

$$\sigma_j = std(\{\mathbf{u}_{ij}^r + \mathbf{f}_i^{r \rightarrow t}\}) \quad 0 \leq i \leq n-1 \ , \tag{5}$$

and select valid flow pseudo labels by a threshold τ. We typically set $\tau = 1$ in our experiments.

After obtaining the valid flow labels from the teacher network, we use them to supervise the student network by a loss function

$$\mathcal{L}_{flow} = \sum_{i=1}^{n} V_i \| (g(\mathbf{I}_i^r, \mathbf{I}^t; \mathbf{W}_t) - g(\mathbf{I}_i^r, \tilde{\mathbf{I}}^t; \mathbf{W}_s)) \| \ , \tag{6}$$

where g is the flow regressor with parameters $\mathbf{W}_t$ and $\mathbf{W}_s$ for the teacher and student network, respectively, and $\tilde{\mathbf{I}}^t$ is the same real image as $\mathbf{I}^t$ but only with stronger data augmentations, and V_i is the mask containing valid pixels whose corresponding 3D keypoint $\mathbf{P}_j$ has $\sigma_j < \tau$. Note that, V_i is generated dynamically from the consistency check between multiple views, and does not rely on any auxiliary information.

3.3. Flow-Guided Photometric Consistency

The previous section only investigates the consistency between the synthetic views and the real input, which potentially suffers from the domain gap problem [44, 19, 53]. To handle this, we add the consistency check between multiple real inputs. Our motivation is the 2D image reprojections on different real images from the same 3D object keypoint should have similar textures. We force the flow network to learn this texture mapping guided by the flow consistency.

Given the real image $\mathbf{I}^t$ with the initial pose $\mathbf{P}_0$ in the previous section, we randomly retrieve another m real images whose initial pose is around $\mathbf{P}_0$, generating m image pairs

$$\{(\mathbf{I}_0^r, \mathbf{I}_k^t)\}, \quad 1 \leq k \leq m, \tag{7}$$

and after feeding the two images in each image pair to the student network, we have

$$\bar{\mathbf{u}}_k^t = \mathbf{u}_0^r + g(\mathbf{I}_0^r, \mathbf{I}_k^t; \mathbf{W}_s) \ , \tag{8}$$

where $\bar{\mathbf{u}}_k^t$ is the predicted 2D image location on image $\mathbf{I}_k^t$. We assume these predicted 2D image locations of the same 3D keypoint have similar texture properties, we add a photometric loss to enforce this

$$\mathcal{L}_{photo} = \sum_{k=1}^{m} V_0 \rho(w(\mathbf{I}_k^t, \bar{\mathbf{u}}_k^t), w(\mathbf{I}_t, \bar{\mathbf{u}}^t)) \ , \tag{9}$$

where w is a operation function that warps the image according to the new pixel locations, ρ is a generalized Charbonnier function to measure the photometric difference based on the Census transformation [36], and $\bar{\mathbf{u}}^t$ is inferred from the student's prediction, where

$$\bar{\mathbf{u}}^t = \mathbf{u}_{0j}^r + g(\mathbf{I}_0^r, \tilde{\mathbf{I}}^t; \mathbf{W}_s) \tag{10}$$

We combine the flow consistency and photometric consistency into our final loss

$$\mathcal{L} = \mathcal{L}_{self} + 0.5\mathcal{L}_{photo}. \tag{11}$$

Note that, we only apply the loss to the student network since the gradient backprogataion only occurs for the student network, and the teacher network only gets its weight updated by EMA updating as discussed in Section 3.1.

4. Experiments

In this section, we first present the experiment setting of our method and then compare our method with state-of-the-art self-supervised methods. We finally conduct detailed ablation studies of our method with different hyper-parameters in different scenarios.

4.1. Experiment Setup

Datasets. We evaluate our method on three widely-used datasets for 6D object pose estimation: LINEMOD ("LM") [11], Occluded-LINEMOD ("LM-O") [21], and YCB-V [54]. The LINEMOD dataset contains 13 objects, with a single sequence per object without occlusions. We follow [51, 13] to use 15% of the real images for training, resulting in a total of 2.4k images. Occluded-LINEMOD is an extension of LINEMOD, which annotates all the objects in one sequence in LINEMOD as the test set and shares the training set with LINEMOD. The recent YCB-V dataset consists of 130k real training images for 21 texture-less objects captured in cluttered scenes. Although all these three datasets contain manually labeled annotations, we train our models on them without accessing the ground truth, and report the final accuracy on their test set. We use the synthetic dataset used in the BOP challenge [6, 12, 46] to train WDR-Pose [16] for the pose initialization.

Evaluation Metrics. We mainly use ADD-0.1d as our metric, which computes the average distance between the mesh vertices transformed by the predicted pose and the ground truth pose, and then only treat the prediction with an average distance below 10% of the mesh diameter as a correct pose estimate. We will use its symmetric version for symmetric objects. Additionally, we also use BOP metrics [12] for evaluation, including the Visible Surface Discrepancy (VSD), the Maximum Symmetry-aware Surface Distance (MSSD), and the Maximum Symmetry-aware Projection Distance (MSPD). We refer the readers to [12] for their detailed definition.

Training details. We use RAFT [49] as our flow regressor for both the teacher and student network. We train the model using AdamW optimizer [32] with a batch size of 16. We use One-cycle strategy [41] to anneal the learning rate from a starting point 4e-4. We crop the target object from the original image according to the initial pose, and then resize the image patch to 256*256. We do not use any data augmentation in the teacher network, and only use the standard random color augmentation used in PFA [13] for the student network. Unlike [51, 50, 28, 17] that train a separate model for each object, which is cumbersome to train, we train a single model for all the objects in a dataset.

4.2. Comparison against State of the Art

We first compare our method against the state-of-the-art self-supervised pose estimation methods on LINEMOD and

Method	DSC [57]	Sock et al. [42]	Lin et al. [28]	Self6D [50]	Self6D‡ [50]	Ours
Ape	31.2	37.6	67.5	76.0	75.4	**81.9**
Bench.	83.0	78.6	**99.9**	91.6	94.9	95.0
Cam	49.6	65.6	87.4	**97.1**	97.0	94.2
Can	56.5	65.6	99.2	**99.8**	99.5	96.8
Cat	57.9	52.5	94.3	85.6	86.6	**95.4**
Driller	73.7	48.8	97.6	98.8	**98.9**	94.8
Duck	31.3	35.1	67.2	56.5	68.3	**83.5**
Eggbox*	96.0	89.2	98.9	91.0	**99.0**	93.9
Glue*	63.4	64.5	96.2	92.2	96.1	**96.5**
Holep.	38.8	41.5	49.9	35.4	41.9	**84.5**
Iron	61.9	80.9	99.5	**99.5**	99.4	94.9
Lamp	64.7	70.7	**99.8**	97.4	98.9	94.8
Phone	54.4	60.5	91.5	91.8	**94.3**	94.1
Avg.	58.6	60.6	88.4	85.6	88.5	**92.2**

Table 1. **Comparison with the self-supervised methods on LINEMOD.** "*" denotes the symmetric objects. Self6D is shorted for Self6D++ [50], which is supervised with RGB images, while Self6D‡ is supervised with additional depth data.

Method	DSC [57]	Sock et al. [42]	Lin et al. [28]	Self6D [50]	Self6D‡ [50]	Ours
Ape	9.1	12.0	40.3	57.7	59.4	**60.1**
Can	21.1	27.5	75.2	95.0	**96.5**	94.2
Cat	26.0	12.0	35.0	52.6	**60.8**	56.5
Driller	33.5	20.5	68.5	90.5	**92.0**	89.7
Duck	12.2	23.0	25.7	26.7	30.6	**30.9**
Eggbox*	39.4	25.1	44.7	45.0	51.1	**58.1**
Glue*	37.0	27.0	60.7	87.1	88.6	**88.9**
Holep.	20.4	35.0	28.0	23.5	38.5	**44.2**
Avg.	24.8	22.8	47.3	59.8	64.7	**65.4**

Table 2. **Comparison with the self-supervised methods on Occluded-LINEMOD.** "*" denotes the symmetric objects. Self6D is shorted for Self6D++ [50], which is supervised with RGB images, while Self6D‡ is supervised with additional depth data.

Occluded-LINEMOD. Since most of them report numbers only in ADD-0.1d, we follow the same for a fair comparison. Table. 1 and 2 summarize the result. Our method outperforms the state-of-the-art methods significantly. Especially, our method, which requires only RGB images, even outperforms Self6D‡ [50] by 3.7% on LINEMOD, which is a method that relies on additional depth images.

4.3. Ablation Study

Evaluation of losses. Table. 3 summarizes the results of our method with different loss combinations. The first row is the results of the standard teacher-student structure used in [43]. Although it already has the EMA updating strategy widely used in self-supervised methods, its performance is still limited without our losses, since the standard teacher-student structure can not effectively select the high-quality pose labels. After adding our flow loss and photometric loss, the performance increases significantly. Combining them achieves the best performance, demonstrating the effectiveness of our method.

$\mathcal{L}_{flow}$	$\mathcal{L}_{photo}$	MSPD	MSSD	VSD	ADD
-	-	0.759	0.589	0.519	30.9
-	✓	0.765	0.631	0.578	37.5
✓	-	0.780	0.711	0.658	64.2
✓	✓	**0.785**	**0.749**	**0.664**	**67.4**

Table 3. **Evaluation of our proposed components.** $\mathcal{L}_{flow}$ is the key component of our framework, while $\mathcal{L}_{photo}$ can improve the performance further.

Ratio	MSPD	MSSD	VSD	ADD
0	0.761	0.605	0.534	41.7
1	0.778	0.743	0.661	61.9
5	0.782	0.744	0.663	63.2
10	0.784	0.745	0.664	65.5
20	0.782	0.748	0.665	65.8
40	0.780	0.749	0.664	66.8
80	0.783	0.748	0.660	67.0
100	**0.785**	**0.749**	**0.665**	**67.4**

Table 4. **Effect of training data size on YCB-V.** With even 1% data, our method can achieve a comparable performance.

Method	MSPD	MSSD	VSD	ADD
Initialization	0.748	0.484	0.367	35.8
Ours + Synthetic	0.871	0.677	0.532	57.0
Ours + Mixed	**0.876**	**0.746**	**0.587**	**66.8**
Ours + SSL	0.859	0.725	0.574	65.4

Table 5. **Effect of real data annotations on LM-O.** "Ours + Synthetic" is trained only on synthetic data, and "Ours + Mixed" is trained on the mixed data containing both the synthetic and real, using both annotations. "Ours + SSL" is our default self-supervised version trained on the real data without annotations.

Evaluation with limited data. We evaluate our method on YCBV with different amounts of real data used in training, and summarize the results in Table. 4. Our method can increase the performance by about 20.2% (from 41.7% to 61.9%) in ADD-0.1d by only using 1% of all the real data. 99% more real data can only increase the performance by 5.5% further, demonstrating the effectiveness of our method in data-limited scenarios.

Effect of real data annotations. In principle, our method can be trained with ground truth pose annotations easily, which is basically training the student network in a stan-

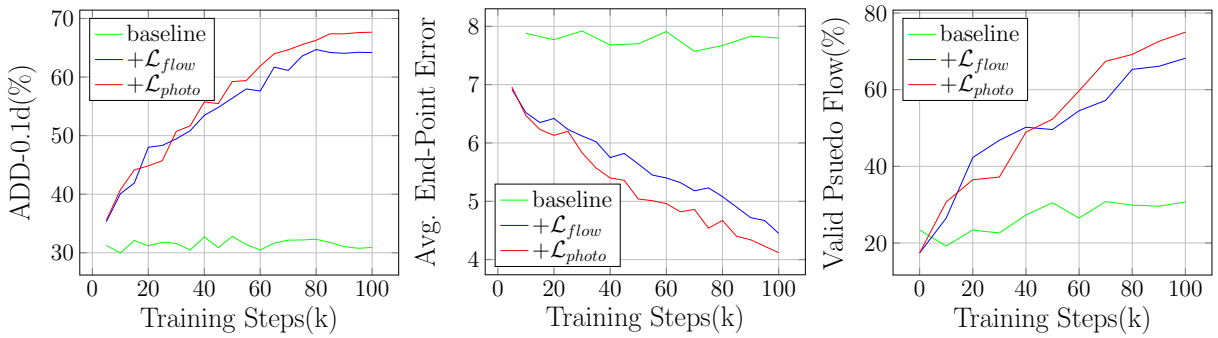


Figure 5. **Training analysis on YCB-V.** We evaluate our method in three different settings during the training. The baseline is the original teacher-student structure proposed in [48], and the other two are the proposed components of our method. The baseline model struggles to learn from the unannotated data due to the lack of proper constraints in selecting high-quality pseudo flow labels. Our flow loss tackles this problem effectively, and the proposed photometric loss increases performance further.

		MSPD	MSSD	VSD	ADD
	2	0.780	0.745	0.662	66.8
	3	**0.785**	**0.749**	**0.664**	67.4
m	4	0.782	0.752	0.661	**67.6**
	5	0.774	0.741	0.657	66.8
	6	0.772	0.736	0.648	65.2
	2	0.776	0.728	0.645	61.6
	3	0.779	0.747	0.667	67.0
n	4	**0.785**	**0.749**	**0.664**	**67.4**
	5	0.782	0.738	0.653	64.7
	6	0.780	0.733	0.649	63.3
	0.5	0.776	0.741	0.655	65.0
	1.	**0.785**	**0.749**	**0.664**	67.4
τ	4.	0.777	0.748	0.662	**67.6**
	8.	0.772	0.736	0.654	65.4
	16.	0.769	0.734	0.647	64.6

Table 6. **Evaluation of hyper-parameters.** n and m are respectively the number of views used for $\mathcal{L}_{flow}$ and $\mathcal{L}_{photo}$, and τ determines the valid prediction for $\mathcal{L}_{flow}$. Our method is robust to the choices of these parameters.

dard fully-supervised way. We compare our self-supervised method with the version trained with all real pose annotations, as shown in Table 5. Our default version, "Ours + SSL" achieves comparable performance as the fully-supervised version and still significantly outperforms the other baselines in Table 2. We show some qualitative results in Fig. 7.

Evaluation of hyper-parameters. We evaluate the hyper-parameters used in our framework in Table 6. We first evaluate the impact of the view numbers used in Section 3.2 and 3.3. It shows more views generally can increase the performance, since it adds more information to the geometry constraint. However, too many views, such as those larger than 4, will make the performance worse. We believe it is caused by the noise introduced by too many views with large viewpoint differences, which usually makes the network harder to learn. We then evaluate the threshold τ used in our framework, which is used to determine the reliability of pseudo flow labels. Generally, The performance suffers when this threshold is too strict or loose. Nevertheless, it works quite stable between 1 and 4.

Training analysis on YCB-V. We evaluate our method in three different settings during the training, as shown in Fig. 5. The baseline is the original teacher-student structure proposed in [48], and the other two are the proposed components of our method. The baseline model struggles to learn from the unannotated data due to the lack of proper constraints in selecting high-quality pseudo flow labels. Our flow loss introduces a constraint based on flow consistency derived from multiview geometry, and tackles this problem effectively, and the proposed photometric loss increases performance further.

Running time analysis. We conduct all our experiments on a workstation with an NVIDIA RTX-3090 GPU and an Intel-Xeon CPU with 12 cores running at 2.1GHz. During inference, our method takes only 23ms for a single object, including the optical flow estimation 17ms and the PnP solver 6ms.

5. Conclusion

We propose a self-supervised 6D object pose method without relying on either depth image or 2D mask annotations. After obtaining a rough pose initialization based on network training on synthetic images, we refine the pose with a teacher-student pseudo labeling framework. To solve the problem of identifying high-quality labels for object pose. We first formulate pseudo object pose labels as pixel-level optical flow supervision signals. Then, we introduce the flow consistency based on the underlying geometry constraint from multiple views. Our experiments show that our method significantly outperforms existing solutions and

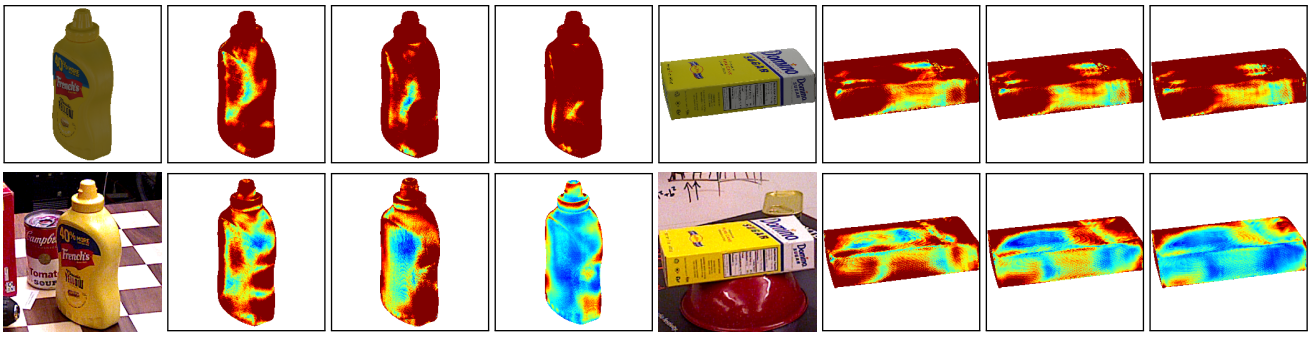


Figure 6. **Visualization of σ across increasing training iterations.** The top row shows the real and synthetic input. We compare our method against the baseline proposed in [48]. As can be seen, the model tends to produce more confident predictions on texture-well regions, where σ is relatively lower. With the aid of our proposed geometry constraint for model training, the model can gradually predict confident 2D matches across the entire surface of the object.

Figure 7. **Qualitative results on YCB-V and LM-O.** We show the results trained only on synthetic data on the top row, and the results fine-tuned using our proposed self-supervised strategy on the bottom row. Our method significantly improve the baseline in various scenarios, such as occluded, weak-textured, and symmetrical objects.

does not rely on other auxiliary information.

References

[1] Dingding Cai, Janne Heikkilä, and Esa Rahtu. Sc6d: Symmetry-agnostic and correspondence-free 6d object pose

estimation. In *International Conference on 3D Vision*, 2022. 2

[2] Bo Chen, Alvaro Parra, Jiewei Cao, Nan Li, and Tat-Jun Chin. End-to-end learnable geometric vision by backpropagating pnp optimization. In *Conference on Computer Vision and Pattern Recognition*, 2020. 2

[3] Hansheng Chen, Pichao Wang, Fan Wang, Wei Tian, Lu Xiong, and Hao Li. Epro-pnp: Generalized end-to-end probabilistic perspective-n-points for monocular object pose estimation. In *Conference on Computer Vision and Pattern Recognition*, 2022. 2

[4] Kai Chen, Rui Cao, Stephen James, Yichuan Li, Yun-Hui Liu, Pieter Abbeel, and Qi Dou. Sim-to-real 6d object pose estimation via iterative self-training for robotic bin picking. In *European Conference on Computer Vision*, 2022. 1, 3, 4

[5] Alvaro Collet, Manuel Martinez, and Siddhartha S Srinivasa. The moped framework: Object recognition and pose estimation for manipulation. *The international journal of robotics research*, 30(10):1284–1306, 2011. 1

[6] Maximilian Denninger, Martin Sundermeyer, Dominik Winkelbauer, Dmitry Olefir, Tomas Hodan, Youssef Zidan, Mohamad Elbadrawy, Markus Knauer, Harinandan Katam, and Ahsan Lodhi. Blenderproc: Reducing the reality gap with photorealistic rendering. In *International Conference on Robotics: Sciene and Systems*, 2020. 2, 5

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

[8] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 2020. 1

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

[10] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Conference on Computer Vision and Pattern Recognition*, 2020. 1

[11] Stefan Hinterstoisser, Vincent Lepetit, Slobodan Ilic, Stefan Holzer, Gary Bradski, Kurt Konolige, and Nassir Navab. Model based training, detection and pose estimation of texture-less 3d objects in heavily cluttered scenes. In *Asian Conference on Computer Vision*, 2012. 2, 5

[12] Tomáš Hodaň, Frank Michel, Eric Brachmann, Wadim Kehl, Anders Glent Buch, Dirk Kraft, Bertram Drost, Joel Vidal, Stephan Ihrke, Xenophon Zabulis, Caner Sahin, Fabian Manhardt, Federico Tombari, Tae-Kyun Kim, Jiří Matas, and Carsten Rother. BOP: Benchmark for 6D object pose estimation. *European Conference on Computer Vision*, 2018. 1, 5

[13] Yinlin Hu, Pascal Fua, and Mathieu Salzmann. Perspective flow aggregation for data-limited 6d object pose estimation. In *European Conference on Computer Vision*, 2022. 1, 2, 3, 4, 5

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

[15] Yinlin Hu, Joachim Hugonot, Pascal Fua, and Mathieu Salzmann. Segmentation-driven 6d object pose estimation. In *Conference on Computer Vision and Pattern Recognition*, 2019. 1, 2

[16] Yinlin Hu, Sebastien Speierer, Wenzel Jakob, Pascal Fua, and Mathieu Salzmann. Wide-depth-range 6d object pose estimation in space. In *Conference on Computer Vision and Pattern Recognition*, 2021. 1, 2, 3, 5

[17] Lin Huang, Tomas Hodan, Lingni Ma, Linguang Zhang, Luan Tran, Christopher Twigg, Po-Chen Wu, Junsong Yuan, Cem Keskin, and Robert Wang. Neural correspondence field for object pose estimation. In *European Conference on Computer Vision*, 2022. 5

[18] Shun Iwase, Xingyu Liu, Rawal Khirodkar, Rio Yokota, and Kris M. Kitani. Repose: Fast 6d object pose refinement via deep texture rendering. In *International Conference on Computer Vision*, 2021. 1, 2, 3

[19] Rico Jonschkowski, Austin Stone, Jonathan T Barron, Ariel Gordon, Kurt Konolige, and Anelia Angelova. What matters in unsupervised optical flow. In *European Conference on Computer Vision*, 2020. 5

[20] Hiroharu Kato, Yoshitaka Ushiku, and Tatsuya Harada. Neural 3d mesh renderer. In *Conference on Computer Vision and Pattern Recognition*, 2018. 2

[21] Alexander Krull, Eric Brachmann, Frank Michel, Michael Ying Yang, Stefan Gumhold, and Carsten Rother. Learning analysis-by-synthesis for 6d pose estimation in rgb-d images. In *International Conference on Computer Vision*, 2015. 2, 5

[22] Donghyeon Kwon and Suha Kwak. Semi-supervised semantic segmentation with error localization network. In *Conference on Computer Vision and Pattern Recognition*, 2022. 3

[23] Y. Labbe, J. Carpentier, M. Aubry, and J. Sivic. Cosypose: Consistent multi-view multi-object 6d pose estimation. In *European Conference on Computer Vision*, 2020. 2, 3

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

[25] Gang Li, Xiang Li, Yujie Wang, Yichao Wu, Ding Liang, and Shanshan Zhang. Pseco: Pseudo labeling and consistency training for semi-supervised object detection. In *European Conference on Computer Vision*, 2022. 3

[26] Yi Li, Gu Wang, Xiangyang Ji, Yu Xiang, and Dieter Fox. Deepim: Deep iterative matching for 6d pose estimation. In *European Conference on Computer Vision*, 2018. 1, 2, 3

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

[28] Haotong Lin, Sida Peng, Zhize Zhou, and Xiaowei Zhou. Learning to estimate object poses without real image anno-

tations. In *International Joint Conference on Artificial Intelligence*, 2022. 1, 3, 4, 5, 6

[29] Lahav Lipson, Zachary Teed, Ankit Goyal, and Jia Deng. Coupled iterative refinement for 6d multi-object pose estimation. In *Conference on Computer Vision and Pattern Recognition*, 2022. 2

[30] Shichen Liu, Tianye Li, Weikai Chen, and Hao Li. Soft rasterizer: A differentiable renderer for image-based 3d reasoning. In *International Conference on Computer Vision*, 2019. 2

[31] Xingyu Liu, Shun Iwase, and Kris M Kitani. Stereobj-1m: Large-scale stereo image dataset for 6d object pose estimation. In *International Conference on Computer Vision*, 2021. 1

[32] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2018. 5

[33] Fabian Manhardt, Wadim Kehl, and Adrien Gaidon. Roi-10d: Monocular lifting of 2d detection to 6d pose and metric shape. In *Conference on Computer Vision and Pattern Recognition*, 2019. 1

[34] Eric Marchand, Hideaki Uchiyama, and Fabien Spindler. Pose estimation for augmented reality: a hands-on survey. *IEEE transactions on visualization and computer graphics*, 22(12):2633–2651, 2015. 1

[35] Pat Marion, Peter R Florence, Lucas Manuelli, and Russ Tedrake. Label fusion: A pipeline for generating ground truth labels for real rgbd data of cluttered scenes. In *International Conference on Robotics and Automation*, 2018. 1

[36] Simon Meister, Junhwa Hur, and Stefan Roth. Unflow: Unsupervised learning of optical flow with a bidirectional census loss. In *AAAI conference on artificial intelligence*, 2018. 5

[37] Nathaniel Merrill, Yuliang Guo, Xingxing Zuo, Xinyu Huang, Stefan Leutenegger, Xi Peng, Liu Ren, and Guoquan Huang. Symmetry and uncertainty-aware object slam for 6dof object pose estimation. In *Conference on Computer Vision and Pattern Recognition*, 2022. 2

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

[39] Sida Peng, Yuan Liu, Qixing Huang, Xiaowei Zhou, and Hujun Bao. Pvnet: Pixel-wise voting network for 6dof pose estimation. In *Conference on Computer Vision and Pattern Recognition*, 2019. 1, 2

[40] Mahdi Rad and Vincent Lepetit. Bb8: A scalable, accurate, robust to partial occlusion method for predicting the 3d poses of challenging objects without using depth. In *International Conference on Computer Vision*, 2017. 2

[41] Leslie N Smith and Nicholay Topin. Super-convergence: Very fast training of neural networks using large learning rates. In *Artificial intelligence and machine learning for multi-domain operations applications*, volume 11006, page 1100612. International Society for Optics and Photonics, 2019. 5

[42] Juil Sock, Guillermo Garcia-Hernando, Anil Armagan, and Tae-Kyun Kim. Introducing pose consistency and warp-alignment for self-supervised 6d object pose estimation in color images. In *International Conference on 3D Vision*, 2020. 1, 2, 6

[43] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in neural information processing systems*, 33:596–608, 2020. 1, 2, 3, 4, 6

[44] Austin Stone, Daniel Maurer, Alper Ayvaci, Anelia Angelova, and Rico Jonschkowski. Smurf: Self-teaching multi-frame unsupervised raft with full-image warping. In *Conference on Computer Vision and Pattern Recognition*, 2021. 5

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

[46] Martin Sundermeyer, Tomas Hodan, Yann Labbe, Gu Wang, Eric Brachmann, Bertram Drost, Carsten Rother, and Jiri Matas. Bop challenge 2022 on detection, segmentation and pose estimation of specific rigid objects. *arXiv preprint arXiv:2302.13075*, 2023. 5

[47] Yihe Tang, Weifeng Chen, Yijun Luo, and Yuting Zhang. Humble teachers teach better students for semi-supervised object detection. In *Conference on Computer Vision and Pattern Recognition*, 2021. 3

[48] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30, 2017. 1, 2, 3, 4, 7, 8

[49] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *European Conference on Computer Vision*, 2020. 5

[50] Gu Wang, Fabian Manhardt, Xingyu Liu, Xiangyang Ji, and Federico Tombari. Occlusion-aware self-supervised monocular 6d object pose estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 1, 2, 5, 6

[51] Gu Wang, Fabian Manhardt, Jianzhun Shao, Xiangyang Ji, Nassir Navab, and Federico Tombari. Self6d: Self-supervised monocular 6d object pose estimation. In *European Conference on Computer Vision*, 2020. 1, 2, 5

[52] Gu Wang, Fabian Manhardt, Federico Tombari, and Xiangyang Ji. Gdr-net: Geometry-guided direct regression network for monocular 6d object pose estimation. In *Conference on Computer Vision and Pattern Recognition*, 2021. 1, 2

[53] Yang Wang, Yi Yang, Zhenheng Yang, Liang Zhao, Peng Wang, and Wei Xu. Occlusion aware unsupervised learning of optical flow. In *Conference on Computer Vision and Pattern Recognition*, 2018. 5

[54] Yu Xiang, Tanner Schmidt, Venkatraman Narayanan, and Dieter Fox. Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes. In *Robotics: Science and Systems Conference*, 2018. 1, 2, 5

[55] Mengde Xu, Zheng Zhang, Han Hu, Jianfeng Wang, Lijuan Wang, Fangyun Wei, Xiang Bai, and Zicheng Liu. End-to-end semi-supervised object detection with soft teacher. In *International Conference on Computer Vision*, 2021. 1, 3

[56] Yan Xu, Kwan-Yee Lin, Guofeng Zhang, Xiaogang Wang, and Hongsheng Li. Rnnpose: Recurrent 6-dof object pose refinement with robust correspondence field estimation and pose optimization. In *Conference on Computer Vision and Pattern Recognition*, 2022. 1, 2

[57] Zongxin Yang, Xin Yu, and Yi Yang. Dsc-posenet: Learning 6dof object pose estimation via dual-scale consistency. In *Conference on Computer Vision and Pattern Recognition*, 2021. 1, 2, 6

[58] Bowen Zhang, Yidong Wang, Wenxin Hou, Hao Wu, Jindong Wang, Manabu Okumura, and Takahiro Shinozaki. Flexmatch: Boosting semi-supervised learning with curriculum pseudo labeling. *Advances in Neural Information Processing Systems*, 34:18408–18419, 2021. 1, 2, 3, 4

[59] Hongyu Zhou, Zheng Ge, Songtao Liu, Weixin Mao, Zeming Li, Haiyan Yu, and Jian Sun. Dense teacher: Dense pseudo-labels for semi-supervised object detection. In *European Conference on Computer Vison*, 2022. 3