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

Yang Hai 1, Rui Song 1, Jiaojiao Li 1, David Ferstl 2, Yinlin Hu 2
1 State Key Laboratory of ISN, Xidian University, 2 MagicLeap

Abstract

Most self-supervised 6D object pose estimation methods 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 estimation method that can be trained with pure RGB images without any auxiliary information. We first obtain a rough pose initialization from networks trained on synthetic images rendered from the target’s 3D mesh. Then, we introduce a refinement strategy leveraging the geometry constraint in synthetic-to-real image pairs from multiple different views. We formulate this geometry constraint 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 state-of-the-art self-supervised methods significantly, with neither 2D annotations nor additional depth images.

1. Introduction

The goal of 6D object pose estimation is to accurately estimate the 3D rotation and 3D translation of a rigid object with respect to the camera, which gives essential information about the world beyond classical 2D understanding and is a fundamental component in many applications, such as robotic manipulation [5], autonomous driving [36], and augmented reality [37].

Recent progress in this field has significantly improved the robustness and accuracy of the model [49, 60, 21, 56, 7, 29, 19, 18]. Most of these approaches, however, rely on a large number of real images with accurate 6D pose annotations. But, compared to classical 2D annotation, these 6D annotations are either very hard to obtain [38, 34] or are prone to contain large labeling errors [16, 11, 58]. Some recent methods propose to use techniques based on image synthesis [15] or self-supervised learning [55, 54] to handle this problem. The main problem with synthetic images is the large domain gap to the real images, making the model’s generalization ability suffers in practice [42, 60]. On the other hand, most self-supervised methods rely on additional information. Some can only work with additional depth images [55, 54, 31] or others need pixel-level annotation of a segmentation mask [46, 61], 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 a rough 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 [47, 62]. Pseudo labeling is widely used in many computer vision tasks [52, 59, 12, 8]. However, the two fundamental problems of pseudo labeling are still open questions in 6D object pose estimation, including the generative strategy of creating pseudo labels and the selection strategy of extracting high-quality labels from the noisy candidates, as shown in Fig. 2.
We propose to formulate the pseudo 6D pose labels as pixel-level flow supervision signals in a render-and-compare framework [16, 26, 29, 60, 21, 32, 9]. Unlike the common render-and-compare frameworks that need accurate pose annotations, we propose a geometry-guided learning framework without any annotations. We render multiple images near the initial pose, and compare them with the real input with the guidance of flow consistency, based on the geometry constraints between these image pairs from different views. We choose high-quality flow labels based on the proposed consistency on the fly, and supervise the network training with these dynamically generated labels in every training step.

We evaluate our method on three challenging datasets LINEMOD [14], Occluded-LINEMOD [24], and YCB-V [58], and show that it outperforms state-of-the-art self-supervised methods significantly, including those methods relied on depth image [55, 54] or auxiliary annotation information [61, 46].

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 any auxiliary information.

2. Related Work

Object pose estimation has shown significant progress recently, based on different techniques, such as direct pose regression [56, 7, 26], 2D reprojection regression [44, 42, 18, 19, 40], 3D keypoint prediction [30, 41, 49, 11], and differentiable PnP solver [27, 17, 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 [58, 16].

Some recent methods tackle this problem by training on synthetic images rendered from the target’s 3D mesh [11, 42, 1], but this strategy suffers from the domain gap between the synthetic and real image sets [6]. In contrast, some pose refinement methods have shown significant improvement in the generalization ability across different domains [29, 32, 26, 21, 60, 9] and especially [16], which produces comparable results as the state-of-the-art methods with only about one-tenth of the real images involved in training. Although having 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 [55, 61, 46, 54] try to remove the cumbersome procedure of pose annotation completely. Most of them are based on a strategy that compares the synthetic image rendered from an initial pose with the real image, and backpropagate the gradient through a differentiable renderer [23, 33] to update the network’s weights during training, expecting to align the rendered image with the real input without explicit annotations. 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, making them often rely either on depth [55] or on additional pixel-level annotations of segmentation masks [61]. In contrast, we propose to compare multiple synthetic-to-real image pairs at the same time, and force networks to comply with the geometry constraint between those image
Figure 3. Method overview. We first obtain the initial pose based on a pose estimation network trained only on synthetic images, and then train our refinement framework on real images without any annotations. Our proposed framework is based on a teacher-student learning scheme. 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 dynamically produce pixel-level flow supervision signals for the student network during the training, by leveraging the geometry-guided flow consistency between those image pairs from different views. After getting 3D-to-2D correspondences based on the predicted flow, we use a PnP solver to get the final pose [16].

Pairs from different views, which suffers little from the domain problem. On the other hand, we formulate the geometry constraint as pixel-level consistency, which provides us dynamic valid label mask during training, without any 3D depth information or additional 2D mask annotations.

Pseudo labeling is also one of the basic techniques used in recent self-supervised object pose methods [31, 4]. However, these approaches still rely on additional depth images to select valid pseudo labels. In addition, they only update the pseudo labels after finishing the previous training, which usually means the model needs to be trained multiple times to utilize the slowly updated pseudo labels. By contrast, our pseudo flow labels are generated dynamically in every training step, and our model only needs to be trained once.

Our method is related to the recent teacher-student formulation of pseudo labeling [47, 62, 52, 59, 25, 64, 28, 63, 57, 43, 22], 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 different data augmentations. Although this simple general framework has been widely used in image classification [47, 62], object detection [59, 28, 64, 51], and semantic segmentation [25], it only can work with high-quality pseudo labels. However, there is still no easy way to generate high-quality 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 flow consistency across multiple different views during training. Our experiments demonstrate the effectiveness of this method.

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 depth images 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 an existing pose estimation network [19, 26] on it to obtain a rough pose initialization [16, 29, 21]. 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 based on the geometry constraint between different views. Finally, we show how we extend it to multiple image pairs to further increase the robustness. Fig. 3 shows the overview of our method.

3.1. Framework Overview

We use a teacher-student architecture [52] 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 different data augmentations. After updating the weights of the student network by gradient backpropagation supervised by this pseudo label, we then update the weights of the teacher network by a simple exponential moving averaging (EMA) strategy from the student network:

$$W_t = \alpha W_t + (1 - \alpha) W_s,$$

where $W_t$ and $W_s$ are the network weight parameters of the teacher and student network, respectively, and $\alpha$ is the exponential factor, which is typically 0.999. The weight updating and pseudo label generation is conducted after each
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 estimation as a problem of estimating dense 2D-to-2D correspondence, or optical flow estimation, as in PFA [16], which, however, is a fully-supervised object pose method. To tackle the problem of no pose annotation for the common correspondence, or optical flow estimation, as in PFA [16], although its true pose $P_t$ is unknown to us, the relation between the 3D keypoint $p_j$ and its 2D image location $u_j^i$ should follow the perspective principle of Eq. 3, implicitly generating the correspondence $p_j \leftrightarrow u_j^i$.

We train a network to predict dense 2D-to-2D correspondence $F_{ij}^{r-t}$ between the two images in each image pair of Eq. 2, such that

$$u_{ij}^r + F_{ij}^{r-t} = u_{ij}^t,$$

where $u_{ij}^t$ is the corresponding 2D flow vector. Although $u_{ij}^t$ is unknown during training, we have the geometry constraint that the 2D image locations $\{u_{ij}^t, f_{ij}^{r-t}\}$ of the same 3D keypoint $p_j$ from different synthetic views $0 \leq i \leq n - 1$ should be the same.

We use the standard variance of the predicted $u_{ij}^t$ from different views to determine if the current pixel’s flow prediction is a valid pseudo label

$$\sigma_j = std(\{u_{ij}^t, f_{ij}^{r-t}\}) \quad 0 \leq i \leq n - 1,$$

and select valid flow pseudo labels by a simple threshold $\tau$. Fig. 7 shows some visualizations of the variance $\sigma$. After obtaining the valid flow labels from the teacher network, we use them to supervise the student network by a loss function

$$L_{flow} = \sum_{i=0}^{n-1} V_i \| g(I_i', T_i'; W_t) - g(I_i', \bar{T}_i'; W_s) \| ,$$

where $g$ is the flow network with parameters $W_t$ and $W_s$ for the teacher and student network, respectively, and $\bar{T}_i'$ iteration during training, making it much more efficient than other pseudo-label-based object pose methods [4, 31], which can only produce pseudo labels after the whole training pipeline and need to train the model multiple times.

Our main problem is how to select high-quality pseudo label candidates from the noisy predictions of the teacher network. For the image classification task as in [52], one can obtain the label 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 [47, 62]. 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.

For an object having $N$ 3D keypoints, the 2D reprojection of a 3D keypoint $p_j$, $1 \leq j \leq N$, under pose $P_t$ can be obtained by

$$\lambda_{ij}^t \begin{bmatrix} u_{ij}^t \\ 1 \end{bmatrix} = K(R_i p_j + t_i),$$

where $\lambda_{ij}^t$ is a scale factor, $u_{ij}^t$ is the 2D image location, $K$ is the intrinsic camera matrix, and $R_i$ and $t_i$ are the rotation and translation of pose $P_t$, respectively. We then establish 3D-to-2D correspondence $p_j \leftrightarrow u_{ij}^t$ under pose $P_t$. For the real image $I'$, although its true pose $P'$ is unknown to us, the relation between the 3D keypoint $p_j$ and its 2D image location $u_j^t$ should follow the perspective principle of Eq. 3, implicitly generating the correspondence $p_j \leftrightarrow u_j^t$.

We train a network to predict dense 2D-to-2D correspondence $F_{ij}^{r-t}$ between the two images in each image pair of Eq. 2, such that

$$u_{ij}^r + F_{ij}^{r-t} = u_{ij}^t,$$

where $u_{ij}^t$ is the corresponding 2D flow vector. Although $u_{ij}^t$ is unknown during training, we have the geometry constraint that the 2D image locations $\{u_{ij}^t, f_{ij}^{r-t}\}$ of the same 3D keypoint $p_j$ from different synthetic views $0 \leq i \leq n - 1$ should be the same.

We use the standard variance of the predicted $u_{ij}^t$ from different views to determine if the current pixel’s flow prediction is a valid pseudo label

$$\sigma_j = std(\{u_{ij}^t, f_{ij}^{r-t}\}) \quad 0 \leq i \leq n - 1,$$

and select valid flow pseudo labels by a simple threshold $\tau$. Fig. 7 shows some visualizations of the variance $\sigma$. After obtaining the valid flow labels from the teacher network, we use them to supervise the student network by a loss function

$$L_{flow} = \sum_{i=0}^{n-1} V_i \| g(I_i', T_i'; W_t) - g(I_i', \bar{T}_i'; W_s) \| ,$$

where $g$ is the flow network with parameters $W_t$ and $W_s$ for the teacher and student network, respectively, and $\bar{T}_i'$

iteration during training, making it much more efficient than other pseudo-label-based object pose methods [4, 31], which can only produce pseudo labels after the whole training pipeline and need to train the model multiple times.

Our main problem is how to select high-quality pseudo label candidates from the noisy predictions of the teacher network. For the image classification task as in [52], one can obtain the label 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 [47, 62]. 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.

For an object having $N$ 3D keypoints, the 2D reprojection of a 3D keypoint $p_j$, $1 \leq j \leq N$, under pose $P_t$ can be obtained by

$$\lambda_{ij}^t \begin{bmatrix} u_{ij}^t \\ 1 \end{bmatrix} = K(R_i p_j + t_i),$$

where $\lambda_{ij}^t$ is a scale factor, $u_{ij}^t$ is the 2D image location, $K$ is the intrinsic camera matrix, and $R_i$ and $t_i$ are the rotation and translation of pose $P_t$, respectively. We then establish 3D-to-2D correspondence $p_j \leftrightarrow u_{ij}^t$ under pose $P_t$. For the real image $I'$, although its true pose $P'$ is unknown to us, the relation between the 3D keypoint $p_j$ and its 2D image location $u_j^t$ should follow the perspective principle of Eq. 3, implicitly generating the correspondence $p_j \leftrightarrow u_j^t$.

We train a network to predict dense 2D-to-2D correspondence $F_{ij}^{r-t}$ between the two images in each image pair of Eq. 2, such that

$$u_{ij}^r + F_{ij}^{r-t} = u_{ij}^t,$$

where $u_{ij}^t$ is the corresponding 2D flow vector. Although $u_{ij}^t$ is unknown during training, we have the geometry constraint that the 2D image locations $\{u_{ij}^t, f_{ij}^{r-t}\}$ of the same 3D keypoint $p_j$ from different synthetic views $0 \leq i \leq n - 1$ should be the same.

We use the standard variance of the predicted $u_{ij}^t$ from different views to determine if the current pixel’s flow prediction is a valid pseudo label

$$\sigma_j = std(\{u_{ij}^t, f_{ij}^{r-t}\}) \quad 0 \leq i \leq n - 1,$$

and select valid flow pseudo labels by a simple threshold $\tau$. Fig. 7 shows some visualizations of the variance $\sigma$. After obtaining the valid flow labels from the teacher network, we use them to supervise the student network by a loss function

$$L_{flow} = \sum_{i=0}^{n-1} V_i \| g(I_i', T_i'; W_t) - g(I_i', \bar{T}_i'; W_s) \| ,$$

where $g$ is the flow network with parameters $W_t$ and $W_s$ for the teacher and student network, respectively, and $\bar{T}_i'$

Figure 4. **Illustration of geometry-guided consistency.** We predict the flow between synthetic and real images. (a) The 2D flow of the same 3D point from different synthetic views to the real input should be consistent. (b) On the other hand, the target 2D locations of the same 3D point on different real inputs should have similar textures.
is the same real image as \( I^t \) but only with different data augmentations, and \( V_i \) is the mask containing valid pixels where \( \sigma_i < \tau \). Note that, \( V_i \) is generated dynamically from the consistency check between multiple image pairs in Eq. 2, and does not rely on any 2D mask annotations.

3.3. Flow-Guided Photometric Consistency

The previous section only investigates the consistency between the synthetic views and the real input. We further explore the consistency between multiple real inputs. Our motivation is that the 2D image reprojections of the same 3D object keypoint on different real images should have similar textures. We formulate this texture assumption as a photometric consistency, as illustrated in Fig. 4(b).

Given the real image \( I^t \) with the initial pose \( P_0 \), as in the previous section, we randomly retrieve another \( m \) real images whose initial pose is around \( P_0 \), generating \( m \) image pairs

\[
\{ (I^t_0, I^t_k) \}, \quad 1 \leq k \leq m, \tag{7}
\]

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

\[
\bar{u}^t_k = u^t_0 + g(I^t_0, I^t_k; W_t), \tag{8}
\]

where \( \bar{u}^t_k \) is the predicted 2D image location on image \( I^t_k \).

Given these predicted 2D image locations of the same 3D keypoint have similar texture properties, and we use a photometric loss to model this

\[
\mathcal{L}_{\text{photo}} = \sum_{k=1}^{m} V_i \rho(w(I^t_k, \bar{u}^t_k), w(I^t, \bar{u}^t)) \tag{9}
\]

where \( w \) is an operation function that warps the image according to the new pixel locations, \( \rho \) is a generalized Charbonnier function to measure the photometric difference based on the Census transformation [39], and \( \bar{u}^t \) is inferred from the student’s prediction, where

\[
\bar{u}^t = u^t_0 + g(I^t_0, \tilde{I}^t; W_s) \tag{10}
\]

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

\[
\mathcal{L} = \mathcal{L}_{\text{flow}} + 0.5 \mathcal{L}_{\text{photo}}. \tag{11}
\]

Note that, we only apply the loss to the student network since the gradient backpropagation 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 in various settings. Our source code is publicly available at https://github.com/YangHai-1218/PseudoFlow.

4.1. Experiment Setup

Datasets. We evaluate our method on three widely-used datasets for 6D object pose estimation: LINEMOD (“LM”) [14], Occluded-LINEMOD (“LM-O”) [24], and YCB-V [58]. LINEMOD dataset contains 13 objects, with a single sequence per object without occlusions. We follow [55, 16] 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, 15, 50] to train WDR-Pose [19] 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 3D error below 10% of the mesh diameter as a correct pose estimate. We use its symmetric version for symmetric objects. Additionally, in some settings, we also use BOP metrics [15] for evaluation, including the Visible Surface Discrepancy (VSD), the Maximum Symmetry-aware Surface Distance (MSSD), the Maximum Symmetry-aware Projection Distance (MSPD), and their average AR. We refer the readers to [15] for their detailed definition.

Training details. We use RAFT [53] as our flow network for both the teacher and student network. We initialize the weights of both teacher and student network with the weights pretrained on synthetic data and train the model using AdamW optimizer [35] with a batch size of 16. We use One-cycle strategy [45] 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 random color augmentation used in PFA [16] for the student network. We typically set \( \tau = 1, m = 3, \) and \( n = 4 \) in our experiments. Unlike [55, 54, 31, 20] that train a separate model for each object, which is cumbersome to train, we train a single model for all objects in the same 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 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. Espe-
Table 3. Evaluation of different components on YCB-V. $L_{flow}$ is the key component of our framework, and $L_{photo}$ improves the performance further.

<table>
<thead>
<tr>
<th>Method</th>
<th>MSPD</th>
<th>MSSD</th>
<th>VSD</th>
<th>ADD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Initialization v1</td>
<td>0.632</td>
<td>0.491</td>
<td>0.420</td>
<td>27.4</td>
</tr>
<tr>
<td>+ Ours (Real)</td>
<td>0.780</td>
<td>0.731</td>
<td>0.673</td>
<td>64.6</td>
</tr>
<tr>
<td>+ Ours (SSL)</td>
<td>0.759</td>
<td>0.722</td>
<td>0.650</td>
<td>63.2</td>
</tr>
<tr>
<td>Initialization v2</td>
<td>0.673</td>
<td>0.580</td>
<td>0.508</td>
<td>36.0</td>
</tr>
<tr>
<td>+ Ours (Real)</td>
<td>0.775</td>
<td>0.722</td>
<td>0.660</td>
<td>65.3</td>
</tr>
<tr>
<td>+ Ours (SSL)</td>
<td>0.764</td>
<td>0.724</td>
<td>0.643</td>
<td>64.2</td>
</tr>
<tr>
<td>Initialization v3</td>
<td>0.694</td>
<td>0.598</td>
<td>0.522</td>
<td>38.6</td>
</tr>
<tr>
<td>+ Ours (Real)</td>
<td>0.803</td>
<td>0.752</td>
<td>0.686</td>
<td>69.2</td>
</tr>
<tr>
<td>+ Ours (SSL)</td>
<td>0.785</td>
<td>0.749</td>
<td>0.664</td>
<td>67.4</td>
</tr>
</tbody>
</table>

Table 4. Performance with different initialization and additional annotations on YCB-V. We evaluate three versions of pose initialization with different accuracy, including the results obtained by the original WDR-Pose (“Initialization v1”), and also two other versions with pre-cropping the detected regions of interest, based on Mask RCNN and RADet, respectively (“v2” and “v3”). Our self-supervised refinement (“SSL”) boosts the initialization accuracy significantly and achieves similar performance as versions trained with fully-annotated real images (“Real”).

Training analysis on YCB-V. We evaluate our method in three different settings during the training, as shown in Fig. 6. The baseline is the original teacher-student structure [52], and the other two are the proposed components of our method. The baseline model struggles to learn from the unannotated data without explicit quality measurement of pseudo 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, as shown in Fig. 7.

Evaluation of hyper-parameters. We evaluate the hyper-parameters used in our framework in Fig. 8. We first evaluate the impact of the number of different views. More views generally increase the performance, since it adds more information to the geometry constraint. However, too many views, such as those larger than 4, has negative impacts on the performance. 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 $\tau$ used in our framework, which is used to determine the reliability of pseudo flow labels. It...
Figure 5. **Qualitative results.** We show the initialization results trained only on synthetic data at the top, and the refinement results after using our self-supervised strategy at the bottom. Our method significantly improves the baseline in various conditions, such as occluded, weak-textured, and symmetry objects.

Figure 6. **Training analysis on YCB-V.** We report the results in three different settings during training. The baseline is the original teacher-student structure [52], and the other two are the proposed components of our method. The baseline model struggles to learn from unannotated data. By contrast, our flow loss tackles this problem effectively, and our photometric loss increases the performance further.

works well between 1 and 4.

**Evaluation with limited real data.** We evaluate our method on YCB-V with different amounts of real data used in training, as shown in Fig. 8(b). Our method increases 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, which demonstrates the effectiveness of our method in data-limited scenarios.

**Evaluation with different initialization and additional annotations.** In principle, our method can be trained with ground truth pose annotations easily, which is basically training the student network in a standard fully-supervised way. We evaluate our self-supervised method with the version trained with real pose annotations. At the same time, to evaluate the robustness of our framework to different pose initialization, we evaluate three versions of pose initialization with different initialization accuracy, including the results obtained by the original WDR-Pose, and two other versions with pre-cropping the detected regions of interest, based on Mask RCNN [13] and RADet [10], respectively. Table 4 summarizes the results. Although the initial poses have different accuracy, our self-supervised refinement framework boosts their performance significantly, and even achieves similar performance as that trained on fully-annotated real images.

**Comparison against standard optical flow methods.** In principle, one can use a self-supervised optical flow method to directly establish dense 2D-to-2D correspondence between the rendered image and the real image input, without any real pose annotations. However, we find that this strategy hardly can work, mainly due to the large domain gap between the rendered and real images, in which case the standard component of photometric comparison in self-supervised flow methods suffers. We evaluate a typical self-supervised optical flow method SMURF [48], and report the results on YCB-V in Table 5. Our self-supervised strategy suffers little from this domain gap problem and produces
much more accurate pose results than SMURF.

**Symmetry handling.** We use the same strategy as in [16, 44] for symmetric objects in obtaining pose initialization, which restricts the range of poses used in training depending on the objects' symmetry type. We do not explicitly handle the symmetry for the refinement, where we build multiple views around the initial pose and use the consistency constraint between them to find the best flow to align different views within a small pose range. Fig. 9 shows two examples with reflectional symmetry and rotational symmetry, respectively. Note how the predicted poses have different 3D axis from the ground truth pose but align the target mesh with the images well in appearance.
**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. The training of our self-supervised framework is only about 20% slower than its fully-supervised version, consuming about 30 and 24 hours in the typical setting on YCB-V, respectively, which is much more efficient than methods relying on multiple times of retraining [31, 4]. For the inference time, our method is the same as its fully-supervised version and takes only 23ms for a single object, including the optical flow estimation 17ms and the PnP solver 6ms.

**5. Conclusion**

We have introduced a simple self-supervised 6D object pose method. After obtaining the rough pose initialization based on a 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 in the context of object pose estimation, we first formulate pseudo object pose labels as pixel-level optical flow supervision signals. Then, we introduce a flow consistency based on the underlying geometry constraint between multiple different views. Our experiments show that the proposed method significantly outperforms existing solutions in both accuracy and efficiency, without relying on any 2D annotations or additional depth images.

**Acknowledgments.** This work was supported by the Natural Science Foundation of China, the Youth Innovation Team of Shaanxi Universities, the Fundamental Research Funds for the Central Universities under Grant JBF220101, and the 111 Project of China under Grant B08038. We thank Thomas Probst, Daniel Koguciuk, Yuan-Ting Lin, and Lin Sun for helpful discussions.

**References**


[52] 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, 6, 7, 8

[53] Zachary Teed and Jia Deng. RAFT: Recurrent All-Pairs Field Transforms for Optical Flow. In European Conference on Computer Vision, 2020. 5


[60] Yan Xu, Kwan-Yee Lin, Guofeng Zhang, Xiaoang 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


