Semi-Supervised 3D Hand-Object Poses Estimation with Interactions in Time

Shaowei Liu[*1] Hanwen Jiang[*1] Jiarui Xu[1] Sifei Liu[2] Xiaolong Wang[1]
[1] UC San Diego [2] NVIDIA

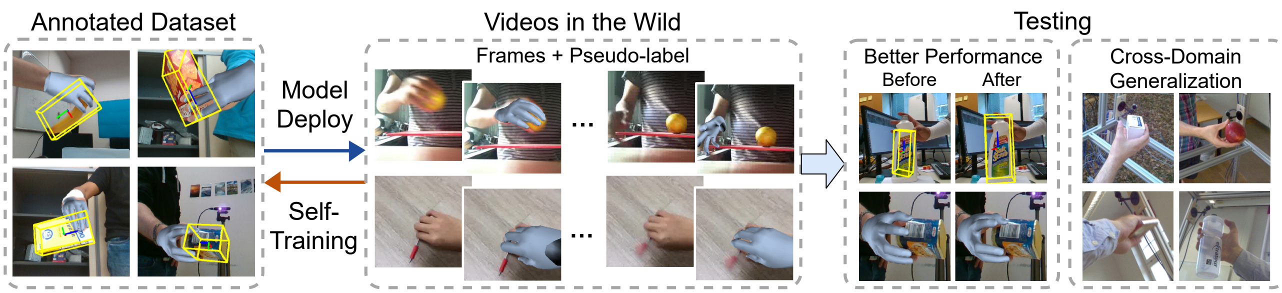


Figure 1: The proposed semi-supervised learning framework and results. We train an initial model on an annotated dataset, and deploy it on a large-scale video dataset to collect pseudo-labels. We leverage spatial-temporal consistency to select pseudo-labels for self-training. After the semi-supervised learning, both the performance of hand-object pose estimation and hand pose generalization are improved.

Abstract

Estimating 3D hand and object pose from a single image is an extremely challenging problem: hands and objects are often self-occluded during interactions, and the 3D annotations are scarce as even humans cannot directly label the ground-truths from a single image perfectly. To tackle these challenges, we propose a unified framework for estimating the 3D hand and object poses with semi-supervised learning. We build a joint learning framework where we perform explicit contextual reasoning between hand and object representations. Going beyond limited 3D annotations in a single image, we leverage the spatial-temporal consistency in large-scale hand-object videos as a constraint for generating pseudo labels in semi-supervised learning. Our method not only improves hand pose estimation in challenging real-world dataset, but also substantially improve the object pose which has fewer ground-truths per instance. By training with large-scale diverse videos, our model also generalizes better across multiple out-of-domain datasets. Project page and code: https://stevenlsw.github.io/Semi-Hand-Object

1. Introduction

Hands are humans' primary means of interacting with the physical world. Capturing the 3D pose of the hands and the objects interacted by hands is a crucial step in understanding human actions. It is also the central part for a variety of applications including augmented reality [44, 25], third-person imitation learning [17, 10], and human-machine interaction [58]. While 3D pose estimation on hands and objects have been studied for a long time in computer vision captured with depth cameras [72, 36, 34, 69, 32] or RGB-D sensors [70, 49, 38, 54, 11] in controlled environments, recent research has also achieved encouraging results on pose estimation from a single monocular RGB image [75, 37, 66].

Despite the efforts, current approaches still highly rely on human annotations for 3D poses, which are extremely difficult to obtain: Researchers have been collecting data with motion capture [50, 13, 74], or aligning mesh models to the real images [19, 30, 16, 4]. Given insufficient annotations for supervised learning, it limits the trained model from generalizing to novel scenes and out-of-domain environments. To enable better estimation performance and generalization ability, we look into video data of hands and objects in the wild, without using the 3D annotations.

Specifically, we propose to exploit hand-object interactions over time. The poses of the hands and objects are usually highly correlated: The 3D pose of the hand when it is grasping the object often indicates the orientation of the object; the object pose also provides constraints on how the hand can approach and interact with the object. When observing from the videos, the 3D poses for both hands and objects should change smoothly and continuously. This continuity provides a cue for selecting coherent and accurate 3D hand and object pose estimation results when human

[*]Equal contribution.

annotations are not available.

In this paper, we introduce a semi-supervised learning approach for 3D hand and object pose estimation with videos. We first train a joint model for both 3D hand pose and 6-Dof object pose estimation with supervised learning using fully annotated data. Then we deploy the model for hand pose estimation in large-scale videos without 3D annotations. We collect the estimation results as novel pseudo-labels for self-training. Specifically, to utilize the interaction information between hand and object, we design a unified framework that extracts the representation from the whole input image, and uses RoIAlign [20] to further obtain the object and hand region representations. Building on these representations, we apply two different branches of sub-networks to estimate the 3D poses for hand and object, respectively. We use a relational module [61] which bridges the two branches for encoding the mutual context between hand and object.

To perform semi-supervised learning with hand-object videos, we deploy our unified model on each frame for pseudo-label generation, as illustrated in Figure 1. Given the 3D hand pose results from our model, we design spatial-temporal consistency constraints to filter unstable and inaccurate estimations. Intuitively, we only keep the results as pseudo-labels if they change continuously over time, which indicates the robustness of the estimation. We then perform self-training with the newly collected data and labels.

We experiment by training the initial model on the HO-3D dataset [16], and perform semi-supervised learning with the Something-Something video dataset [14]. By learning from the pseudo-labels from large-scale videos using our approach, we achieve a large gain over state-of-the-art approaches in the HO-3D benchmark. We also show significant improvements in 3D hand pose estimation which generalizes to the out-of-domain datasets including FPHA [11] and Frei-Hand [76] datasets. More surprisingly, even though we only use pseudo-labels for hands, our joint self-training improves the object pose estimation by a large margin (more than 10% in some objects).

Our contributions include: (i) An unified framework for joint 3D hand and object pose estimation; (ii) A semi-supervised learning pipeline which exploits large-scale unlabeled hand-object interaction videos; (iii) Substantial performance improvement on hand and object pose estimation, and generalization to out-of-domain data.

2. Related Work

Hand pose estimation. Research in RGB-based hand pose estimation can be generally categorized into two paradigms, model-free approaches [5, 75, 26, 57, 37, 12] and model-based approaches [2, 4, 30, 73, 19]. Model-free approaches estimate pose by learning joint coordinates [75, 57] or joint heatmaps [26, 37, 5]. For example, Zimmermann *et al.* [75] proposed to detect 2D hand joints and lift them into 3D with the articulation prior. Differently, model-based approaches utilize the differentiable MANO model [50] to capture hand pose and shape. For example, Boukhayma *et al.* [4] collected synthetic data for pre-training to increase the hand pose accuracy. In our work, instead of relying on either synthetic data or 3D ground-truths, we leverage the spatial-temporal information in the large-scale real-world videos to achieve better hand pose estimation performance and generalization ability in a semi-supervised manner.

Object pose estimation. There are also two main paradigms to perform object 6-Dof pose estimation, with one directly regressing the pose as network outputs [28, 67] and another regressing the projected 3D object control points location in the image and recovering the pose with 2D-to-3D correspondence [45, 60, 43, 24]. Due to the non-linearity of the rotation space, direct regression of the 6-D pose suffers from the generalization problem [24, 65]. For the 2D-to-3D genre, as an example, Hu *et al.* [24] performed inference by generating pixel-wise structural outputs, containing multiple proposals for computing the pose, which demonstrates strong robustness to unseen data. Our method is inspired by this approach, but further extends to consider the hand-object interaction for object pose estimation. We propose a joint framework using contextual reasoning to improve both hand and object pose estimations.

Hand-object interaction. Simultaneously estimating hand and object poses [59, 8, 40, 41, 18, 9, 7] during interaction is a challenging task due to self-occlusion. To tackle this problem, Hasson *et al.* [19] leveraged physical constraints for regressing hand and object mesh at the same time using two separate networks. Differently, we observe that sharing the feature backbone in learning between two pose estimation tasks can implicitly encode the context information. And this contextual information becomes very useful when applied to the semi-supervised learning setting. Similar to our approach, Chen *et al.* [7] proposed to fuse hand-object representations to get interaction-aware features for joint pose estimation using LSTM [15]. However, this feature fusion method cannot model the spatial dependency between hand and object. Instead, our method utilizes a spatial contextual reasoning module between hand-object representations to get interaction-aware feature maps explicitly, which benefits both hand and object pose estimation.

Semi-supervised Learning. Semi-supervised learning plays a key role in improving model performance when the labeled data is limited [46, 35, 68, 55, 71, 48, 51, 6]. Given a trained model on human-annotated datasets, we can apply it on unlabeled data to collect pseudo-labels for further training [31, 27, 1, 56]. For example, Hinton *et al.* [22] have proposed to perform model ensembles in testing to improve the estimation performance. Instead of using multiple models, Radosavovic *et al.* [46] proposed to deploy the trained model with test-time augmentations to increase

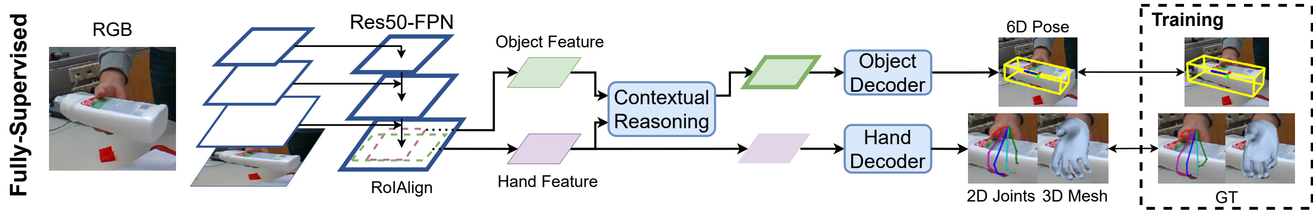


Figure 2: Overview of the framework for hand-object 3D poses joint estimation. The model consists of the shared encoder, the contextual reasoning module for modeling hand-object interaction and two decoders for estimating hand-object poses.

the confidence of the pseudo-labels. While related to our method, most of the previous approaches have not considered the spatial and temporal constraints in videos for selecting the pseudo-labels, which is one of our innovations.

Interaction Reasoning. Interaction reasoning [61, 53, 63, 62, 64, 52, 3, 23] aims to model the interactions among objects. For example, Santoro *et al.* [53] inferred relations across all pairs of objects to solve the visual question answering task. Wang *et al.* [61] captured long-range dependencies via the non-local module in spacetime for video classification. The goal of our work is to exploit the visual correlation between hand and object to improve pose estimation performance under occlusion. The contextual reasoning module we proposed exploits only the relevant pair of cells between hand-object instead of the whole image.

3. Overview

Our method on 3D hand and object pose estimation contains two main components: (i) a joint learning framework with contextual reasoning between the hand and the object; (ii) a semi-supervised pipeline which explores extra labels in videos for training.

First, we present the hand-object joint learning framework in Section 4. The model contains a shared encoder and two separate decoders for hand and object, as well as a contextual reasoning module 4.1 to better exploit their relations. The model is trained under fully-supervised learning.

Then, we propose the semi-supervised learning pipeline in Section 5, Constrained by the spatial-temporal consistency, we generate high-quality pseudo-labels of hand on a large-scale video dataset [14] and re-train our model on the union of fully annotated dataset [16] and those confident pseudo-labels. Because of the diversity in the hand pseudo-labels, the model could both increase the accuracy of hand pose estimation and generalization. With better hand features as context via the contextual reasoning module, the object pose performance of the model could also be improved.

4. Hand-Object Joint Learning Framework

Our hand-object joint learning framework is presented in Figure 2. We use FPN [33] with ResNet-50 [21] as the backbone and extract hand and object features $\mathcal{F}^h$ and $\mathcal{F}^o$ into $\mathbb{R}^{H\times W\times C}$ with the RoIAlign [20], given their corresponding bounding boxes. Then we apply the contextual reasoning between the two features and send the enhanced feature maps with strengthened interactive context information into the hand and the object decoders respectively, which output the 3D hand mesh and 6-Dof object pose. The total loss function of the system is the sum from two decoder branches $\mathcal{L} = \mathcal{L}_{hand} + \mathcal{L}_{object}$. The contextual reasoning module, hand and object decoders, and training losses $\mathcal{L}_{hand}, \mathcal{L}_{object}$ will be discussed in the following sections.

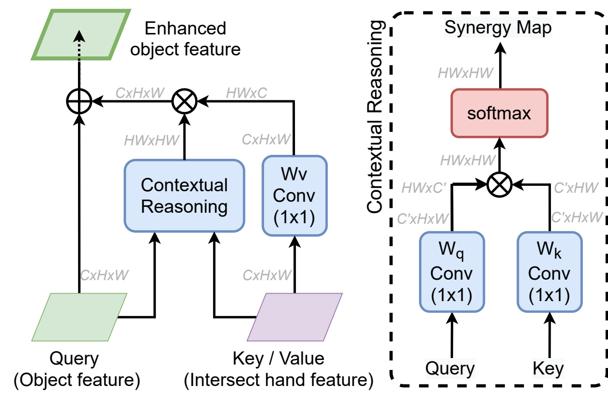


Figure 3: The contextual reasoning (CR) module: The key (features extracted from the region of hand intersecting with object) is taken as context to enhance the query (object features). First, we compute a synergy map as the spatial similarity between query and key. Then we use this map for fusing the key into the query.

4.1. Contextual Reasoning

We exploit the synergy between hand and object features via the contextual reasoning (CR) module as shown in Figure 3, where the query positions in object features could be enhanced by fusing information from the interaction region.

In the module, we take object features $\mathcal{F}^o$ as query and hand-object intersecting regions $\mathcal{F}^{ho}$ as key to model their pairwise relations on the top of RoIAlign [20]. With $\mathcal{F}^{o^+}$ representing the output enhanced object features and subscript i representing the given position i of the features, we have

$$\mathcal{F}_i^{o^+} = \sum_{j\in\Omega} w(i,j)\cdot V(\mathcal{F}_j^{ho}) + \mathcal{F}_i^o \;, \tag{1}$$

where Ω denotes the set of all positions on the key with size

$H \times W$. V is the value transformation function parameterized by W_v shown in the left side of Figure 3, $w(i,j)$ is the pairwise spatial similarity score between query position i and key position j. $w(i,j)$ is computed as

$$w(i,j) = \frac{\exp(q_i^T k_j)}{\sum_{s \in \Omega} \exp(q_i^T k_s)} , \tag{2}$$

where $q_i = W_q F_i^o$ and $k_j = W_k F_j^{ho}$ denote the query and key embedding at position i and j respectively. W_q and W_k, shown in the right side of Figure 3, are the parameters of two separate 1-D convolution that applied on the query and key to get the embeddings.

Besides, we also ablate on using different features as the query of CR module in Section 6.5.1, where we alternatively choose to use the features of hand or both hand-object as the query.

4.2. Hand Decoder

The hand decoder consists of a 2D joints localization network and a mesh regression network. The 2D joints localization network is an hourglass [39] module which takes the hand features after RoIAlign [20] as input and outputs 2D heatmaps for each joint $j \in \mathcal{J}^{2D}$, where $\mathcal{J}^{2D} \in \mathbb{R}^{N^h \times 2}$ and $N^h = 21$ is the number of joints. The heatmaps have the resolution 32×32. The loss function of the 2D joints localization network $\mathcal{L}_{\mathcal{H}}$ is the distance between ground-truth heatmaps $\mathcal{H}_j$ and predictions $\hat{\mathcal{H}}_j$ of each joint j, as $\mathcal{L}_{\mathcal{H}} = \sum_{j \in N^h} ||\mathcal{H}_j - \hat{\mathcal{H}}_j||_2^2$.

The mesh regression network combines the hand features with the 2D heatmaps as input and predicts the parameters of the hand mesh parameterized by the MANO model [50]. The MANO model maps the pose parameters $\theta \in \mathbb{R}^{48}$ and shape parameters $\beta \in \mathbb{R}^{10}$ to hand mesh vertices $\mathcal{V} \in \mathbb{R}^{778 \times 3}$ and 3D joints $\mathcal{J}^{3D} \in \mathbb{R}^{N^h \times 3}$. The inputs of the mesh regression network are forwarded to a ConvNet with four residual blocks [21] and vectorized into a 2048-D feature vector. The output are the predicted MANO parameters $\hat{\theta}$ and $\hat{\beta}$ using three fully-connected layers. We compute the loss on both the MANO model parameters and outputs. Specifically, we compute the L_2 distance between the prediction $(\hat{\theta}, \hat{\beta}, \hat{\mathcal{J}^{3D}}, \hat{\mathcal{V}})$ and ground-truth $(\theta, \beta, \mathcal{J}^{3D}, \mathcal{V})$ as the loss $L_{\mathcal{M}}$ in mesh regression network. The total loss of the hand decoder is the sum of heatmap loss $\mathcal{L}_{\mathcal{H}}$ and $L_{\mathcal{M}}$

$$\mathcal{L}_{hand} = \lambda_{\mathcal{H}} \cdot \mathcal{L}_{\mathcal{H}} + \mathcal{L}_{\mathcal{M}} . \tag{3}$$

where $\lambda_{\mathcal{H}} = 0.1$ is used for balancing losses.

4.3. Object Decoder

The object decoder consists of two streams, which has 4 shared convolution layers and 2 separate convolution layers for each stream. The first stream predicts the 2D location of pre-defined 3D control points on the object from image grid proposals, and the second stream regresses the corresponding confidence scores of each proposal. After obtaining the 2D positions of control points, the object 6-Dof pose can be computed by the PnP algorithm using the correspondence between 2D control points and original 3D control points on the object mesh. In this work, we utilize $N^o = 21$ control points, including 8 corners, 12 edge midpoints and 1 center-point of the object mesh 3D bounding box.

Figure 4: Examples of hand mesh pseudo-labels generated from the video dataset [14].

In the first stream, we adopt the grid-based method [47] to better handle self-occlusion, where each grid g in the object feature map gives a prediction for every control point $i \in N^o$. We use $\delta_{g,i}$ to denote the geometric distance between the grid prediction and the target control point. The loss function of the first stream is the loss sum over all the grids g and control points i, denote as $\mathcal{L}_{p2d} = \sum_g \sum_{i=1}^{N^o} ||\delta_{g,i}||_1$.

The second stream regresses a confidence score $c_{g,i}$ for each grid g and control point i, where the confidence ground-truth $c_{g,i} = \exp(-||\delta_{g,i}||_2)$, which indicates the proximity of the prediction to the ground-truth 2D point locations. During test time, we pick 10 most confident proposals as the input of the PnP algorithm to solve for the object pose. The loss function of the second stream is denoted as $\mathcal{L}_{conf} = \sum_g \sum_{i=1}^{N^o} ||\hat{c}_{g,i} - c_{g,i}||_2^2$, where $c_{g,i}$ and $\hat{c}_{g,i}$ are the ground-truth and predictions.

The total loss of the object decoder is

$$\mathcal{L}_{object} = \lambda_p \cdot \mathcal{L}_{p2d} + \lambda_c \cdot \mathcal{L}_{conf} . \tag{4}$$

where $\lambda_p = 0.5$ and $\lambda_c = 0.1$ are hyperparameters.

5. Semi-Supervised Learning

After we trained the model of hand-object pose estimation on the fully annotated dataset, we deploy it on a large-scale unlabeled video dataset [14] for 3D hand pseudo-label generation. We leverage spatial-temporal consistency to filter out unreliable pseudo hand labels. The obtained pseudo-label examples on the video dataset [14] can be seen in Figure 4.

Note that we do not generate pseudo-labels for objects because of the need for object 3D models at inference time and the poor generalization of object pose due to the limited instances on the annotated dataset. By enlarging the fully annotated dataset with the selected pseudo-labels, we conduct self-training for both hand and object pose estimation.

5.1. Pseudo-Label Generation

We first deploy our model on video frames from a large-scale video dataset [14] for 3D hand pose estimation. To improve the estimation robustness, we do test-time data augmentation and ensemble the predictions similar to [46]. In our experiment, we perform 8 different augmentations of each instance and average the results. The outputs of each frame include 2D joints $\mathcal{J}^{2D}$, 3D joints $\mathcal{J}^{3D}$, 3D hand mesh vertices $\mathcal{V}$, and corresponding MANO parameters (θ, β).

While ensemble predictions reduce the noise in generated samples, we still need to identify confident ones. To this end, we establish a pipeline for filtering by innovatively utilizing the spatial and temporal consistency constraints in the video dataset, as shown in Figure 5.

5.1.1 Spatial Consistency Constraints

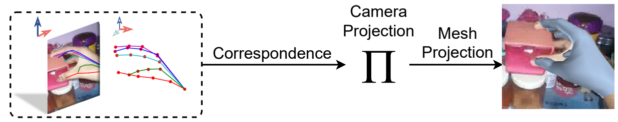
Filtering with spatial consistency requires the corresponding camera pose of each frame. However, it is infeasible to infer the camera pose directly on the video dataset like [14] which has a large variety of viewpoint changes. Our solution to this problem is to leverage the correspondence between the estimated 3D joints $\mathcal{J}^{3D}$ and 2D joints $\mathcal{J}^{2D}$ and solve for the optimal camera parameters Π that projects the 3D joints to 2D, as shown in Figure 5a. We use the weak-perspective camera model and use the SMPLify [42] for the optimization. The objective is the following:

$$\Pi^* = \arg\min_{\Pi} ||\Pi\mathcal{J}^{3D} - \mathcal{J}^{2D}||_2^2 \,, \tag{5}$$

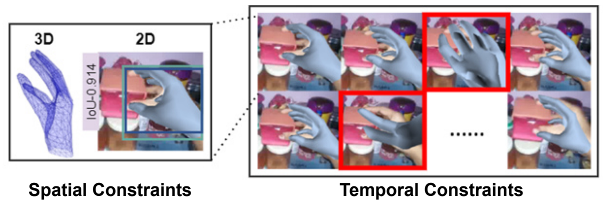
where Π^* is the optimal camera parameters.

IoU Constraint. With the camera pose, we can re-project the estimated 3D mesh $\mathcal{V}$ to the image plane and calculate the Intersection-over-Union (IoU) between the provided ground bounding box B_g and the re-projected mesh bounding box B_d, as shown in the left side of Figure 5b. Note that although we do not have 3D ground-truths, we leverage the 2D bounding box annotations provided by [14]. A confident prediction should always tend to be consistent between these two boxes and has a large IoU. We set the IoU threshold as 0.6 for confident prediction.

Pose Re-projection Constraint. The re-projected 3D hand joints $\Pi\mathcal{J}^{3D}$ and estimated 2D hand joints $\mathcal{J}^{2D}$ should be consistent. We first normalize these two set of joints independent of input sizes and compute the L_2 distance between them as $||\mathcal{J}^{2D} - \Pi\mathcal{J}^{3D}||_2$. When the distance is larger than the threshold t_p, then the prediction would be filtered out. We set $t_p = 0.65$



(a) Solve the camera projection by utilizing correspondence between estimated 3D hand joints and 2D hand joints. Then project the predicted 3D mesh to 2D.



(b) Leverage spatial-temporal consistency for pseudo-label selection. First apply spatial constraints in both 2D and 3D, then perform temporal filtering between consecutive frames.

Figure 5: Pipeline of pseudo-label selection for video frames in the wild.

Biomedical Constraint. The predicted hand pose should be natural human hands. Thus, we exploit the minimal normalized bone length of 0.1 and physically plausible joint angle ranges within $(0, 90)$ as two additional constraints to help remove those unnatural predicted hand poses.

For each instance in the dataset, if it does not violate the three spatial consistency constraints mentioned above, we move forward to the temporal consistency constraints.

5.1.2 Temporal Consistency Constraints

Smoothness Constraint. We consider the temporal consistency constraints as the smoothness of both the 2D joint predictions and the 3D mesh predictions between two consecutive frames $t-1$ and t. As shown in the right side of Figure 5b, since the frames are continuous, the model outputs should be smooth over time. Concretely, the distance of the 2D pose estimation results between these two frames $||\mathcal{J}_t^{2D} - \mathcal{J}_{t-1}^{2D}||_2$ should be less than a threshold t_j. Similarly, for the MANO pose parameter θ, we have $||\theta_t - \theta_{t-1}||_2 \leq t_\theta$ to ensure 3D mesh smoothness, where $t_j = 0.5$ and $t_\theta = 0.01$ are two constant thresholds.

Shape Constraint. In each video sequence, the shape of the hands belongs to the same person should be invariant over time. Given the confident prediction subset C which is the collection of frames satisfying the above constraints in each video sequence, we compute the mean hand shape as $\bar{\beta} = \frac{1}{|C|}\sum_{t\in C}\beta_t$ and its standard deviation $\sigma_C = \sqrt{\frac{1}{|C|}\sum_{t\in C}||\beta_t - \bar{\beta}||_2^2)}$. We filter out the frames in C whose shape deviation from $\bar{\beta}$ is 2 times larger than σ_C.

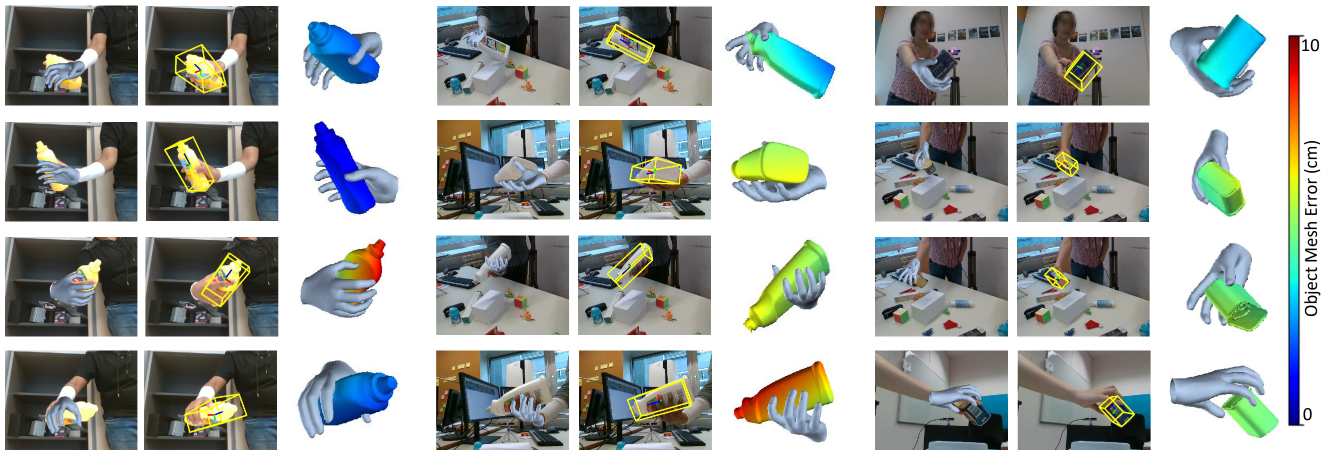


Figure 6: Qualitative results of predicted hand-object pose estimation on the HO-3D dataset. For each example, the first two columns show the recovered hand mesh and estimated 6-Dof object pose, the third row shows the estimated hand-object in 3D, the color of the object indicates the 3D object mesh error. Red indicated larger estimation error.

5.2. Re-training with Pseudo-Labels

We conduct self-training on the union set of the human-annotated dataset and those pseudo-labels. The diversity of the hand pseudo-labels not only improves the hand prediction, but also provides a richer context for hand-object interaction reasoning via the CR module, leading to better object pose estimation. During the retraining, since we do not have the pseudo-labels of objects, we use a binary mask to ensure only computing the loss of the hand on the pseudo-labeled dataset. The total loss function in the retraining stage is the following

$$\mathcal{L} = \mathcal{L}_{hand} + \mathcal{B} \cdot \mathcal{L}_{object} \ . \tag{6}$$

where $\mathcal{B}$ is the mask which equals 1 on the fully-annotated dataset and 0 otherwise.

6. Experiment

First, we test hand-object pose estimation performance on the HO-3D [16] dataset and visualize the prediction in Section 6.4. In Section 6.5, we conduct abundant ablation studies on the designs of the CR module and explore the effectiveness of semi-supervised learning. In section 6.6, we test our model's generalization before and after semi-supervised learning on Freihand [76] and FPHA [11] dataset.

6.1. Implementation Details

Our model is trained in an end-to-end manner from scratch both in the supervised learning phase and semi-supervised learning phase. The shared encoder in the joint learning framework is initialized with ResNet-50 pre-trained on ImageNet. We use a batch size of 24, initial learning rate $1e^{-4}$, and Adam optimizer for the training. The training lasts for 50 epochs and the learning rate is scaled by a factor of 0.7 every 10 epochs. We crop the input image to 512×512 and do data augmentation including scaling ($\pm 20\%$), rotation ($\pm 180°$), translation ($\pm 10\%$) and color jittering ($\pm 10\%$).

6.2. Datasets

HO-3D Dataset [16] is used to train our model in supervised learning. The dataset contains more than 65 sequences captured with 10 different subjects and 10 objects with both 3D pose annotations of hand and object. It has $66,034$ and $11,524$ hand-object interaction images from a third-person view for training and testing. The test results are evaluated by the official online submission system.

Something-Something Dataset [14] is a large-scale hand-object interaction video dataset where we conduct semi-supervised learning with only hand and object bounding boxes provided. It covers a variety of hand instances and most daily objects. We finally select 84063 frames with pseudo hand labels.

FPHA, Freihand Dataset [11, 76] are used for validating cross-domain generalization. FPHA is an egocentric video dataset with hand-object interactions. The dataset is captured by using magnetic sensors strapped on hands. The first-person viewpoint and sensors on the hand introduce great challenges for the model's generalization. We use the same subset as previous work [59, 19, 18] for testing. The test set of the Freihand dataset contains 3960 samples with hands in both indoor and outdoor environments. Hands in half of the samples are interacting with the objects, while in the other half objects are absent. It is also a challenging benchmark for models trained only on hand-object interactions. The evaluation is also performed at the online server.

6.3. Evaluation Metrics

For hand pose estimation, We report the standard metric returned from the submission system, i.e. mean joint error and mesh error in mm after Procrustes alignment and

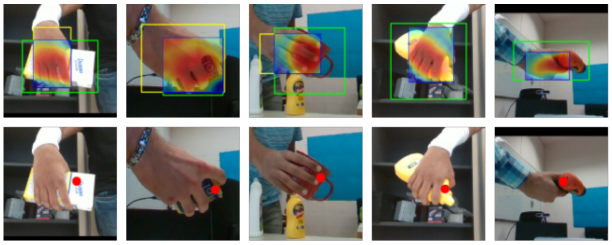


Figure 7: Visualization of the CR module. The top row shows the synergy maps (red indicates higher response), the bottom row shows corresponding object query positions (red points). Hand, object and intersection boxes are in yellow, green, and blue.

methods	Hand Error(↓) Joint	Mesh	F-score(↑) F@5	F@15	Object Estimation
Hasson *et al.* [19]	11.1	11.0	46.0	93.0	✓
Hampali *et al.* [16]	10.7	10.6	50.6	94.2	✗
Ours	**9.8**	**9.4**	**53.0**	**95.7**	✓

Table 1: Hand pose estimation performance compared with state-of-the-art methods on HO-3D [16] dataset. The joint and mesh errors are in mm. The checkmark denotes whether the method also estimates the object pose.

F-scores. To further evaluate cross-dataset generalization, we report the area under the curve (AUC) of the percentage of correct vertices (PCV) and keypoints (PCK) follow [76]. For object pose estimation, we report the percentage of average object 3D vertices error within 10% of object diameter (ADD-0.1D).

6.4. Pose Estimation Performance

Qualitative Results. The qualitative results on the HO-3D dataset [16] are shown in Figure 6. We visualize the predicted hand-object in both 2D and 3D. It demonstrates our method can well handle the occlusion in the interaction and recover the accurate 3D hand mesh and object 6-Dof pose. Moreover, we visualize the synergy maps of different object query positions in the CR module in Figure 7. We observe the CR module gives high responses to contact regions and tends to use the contact pattern for relational reasoning.

Comparison with State-of-the-Art. We compare our hand pose estimation results with state-of-the-art-methods [19, 16] on the HO-3D dataset [16] as shown in the Table 1 and Figure 8. As can be seen from the figure, our method achieves the highest mesh-AUC at 81.2%, 2.2%, and 3.9% higher than [16] and [19] respectively. Our model also has the lowest hand mesh error of 9.4mm. Besides the better hand pose performance, compared with[19, 16], our method could also give a much accurate object estimation simultaneously.

6.5. Ablation Study

We perform the ablation study on HO-3D dataset [16]. We first compare different designs of the CR module. Then, we investigate the effect of different filtering constraints and fractions of pseudo-labels contributed to hand and object pose estimation respectively.

model	Hand Error(↓) Joint	Mesh	F-score(↑) F@5	F@15	Object ADD-0.1D(↑) cleanser	bottle	can	ave
w/o CR	10.2	9.7	**53.7**	94.9	75.3	59.1	52.1	62.2
$\mathbf{h}^+$	10.6	10.1	52.6	94.5	75.5	62.0	**57.6**	65.0
$\mathbf{h}^+\mathbf{o}^+$	10.3	9.9	52.6	94.8	86.5	**61.6**	46.4	64.8
$\mathbf{o}^+$	**10.1**	**9.7**	52.9	**95.2**	**87.6**	60.1	54.8	**67.5**

Table 2: Comparison of different queries in CR module on HO-3D dataset [16]. $\mathbf{h}^+$, $\mathbf{h}^+\mathbf{o}^+$ and $\mathbf{o}^+$ are to take hand features, both hand-object features, and object features as the query respectively. The key feature, which is extracted from the hand-object intersection region, is fused into the query feature. In detail, for $\mathbf{h}^+\mathbf{o}^+$, we use two separate CR modules for fusing the key into both the hand and object query. Ave means average.

model	Hand Error(↓) Joint	Mesh	F-score(↑) F@5	F@15	Object ADD-0.1D(↑) cleanser	bottle	can	ave
sup-w/o CR	10.2	9.7	**53.7**	94.9	75.3	59.1	52.1	62.2
semi-w/o CR	10.1	9.6	**53.7**	95.2	88.8	68.9	49.8	69.2
sup- w/ CR	**10.1**	**9.7**	52.9	**95.2**	**87.6**	**60.1**	**54.8**	**67.5**
semi- w/ CR	**9.8**	**9.4**	53.0	**95.7**	**89.7**	**72.7**	**57.0**	**73.2**

Table 3: Ablation analysis of semi-supervised learning on HO-3D dataset [16]. **sup** and **semi** means under the supervised and semi-supervised phase. Best numbers of supervised learning results and semi-supervised learning results are shown in blue and red respectively. **w/o CR** is the baseline without CR module, and **w/ CR** is the proposed one with CR module. Ave means average.

6.5.1 Ablation on CR Module

We study the effect of different query design choices in the CR module under supervised learning. We compare three types of queries, while the key remains the same as the feature from the intersection region between hand-object. The first choice is to take the hand features as query and the CR module output is only fed into the hand decoder, denoted as $\mathbf{h}^+$. The second choice is to take both the hand and object features as queries using two separate CR modules and the output is fed into both decoders, denoted as $\mathbf{h}^+\mathbf{o}^+$. While the third one is what we proposed in Section 4.1 where we take object features as query, denoted as $\mathbf{o}^+$. As shown in Table 2, performing contextual reasoning to enhance object representation ($\mathbf{o}^+$) can improve the object pose estimation significantly. The average object ADD-0.1D has a 5.3% improvement against the baseline without the CR module. However, using the hand features as a query ($\mathbf{h}^+$ and $\mathbf{h}^+\mathbf{o}^+$) does not contribute to better hand pose estimation. It even degrades the performance slightly against the baseline. This might be due to the CR module degenerates and the occlusion regions even distract the network attention.

6.5.2 Ablation on Semi-supervised Learning

Semi-supervised Learning Performance. As shown in Table 3, semi-supervised learning improves both hand and

model	Num Pseudo	Hand Error(↓) Joint	Mesh	F-score(↑) F@5	F@15
sup	0	10.1	9.7	52.9	95.2
w/o spatial	84,063	10.0	9.7	51.8	95.4
w/o temporal	84,063	10.1	9.7	52.0	95.3
proposed	84,063	**9.8**	**9.4**	**53.0**	**95.7**

Table 4: Ablation analysis of using different filtering constraints in semi-supervised learning on the HO-3D dataset.

eval / model	FPHA [19] AUC (↑)		F-Score (↑)		Freihand [76] AUC (↑)		F-Score (↑)	
	Joint	Mesh	F@5	F@15	Joint	Mesh	F@5	F@15
supervised	73.9	75.0	43.8	92.0	68.5	68.0	34.2	84.8
semi-supervised	**74.8**	**75.9**	**45.3**	**92.9**	**69.7**	**69.1**	**35.4**	**86.3**

Table 5: Cross-dataset generalization performance measured as joint AUC, mesh AUC, and F-scores between model trained with supervised learning and semi-supervised learning.

object pose estimation with or without using the CR module. Even though only hand pseudo-label are collected, the object pose estimation is benefited. We conjecture there are two reasons: First, better hand representation acts as a better context and could improve the object pose estimation via contextual reasoning; Second, the diversity of the hand pseudo-hand labels contribute to a much more robust shared backbone, which extracts base features for obtaining both hand and object representations via RoI Align. With semi-supervised learning, the object ADD-0.1D of cleanser, bottle, and can are improved by 2.1%, 12.6%, 2.2% respectively, as a 5.7% improvement in average.

Pseudo-labels Filtering Constraints. We evaluate how the different filtering constraints contribute to the semi-supervised learning in Table 4. We compare against the method with only supervised learning and methods that remove the spatial or temporal filtering constraints. From the table, We can see each constraint plays an important role. Without either spatial or temporal constraints, it even degrades the hand pose estimation accuracy. Therefore, both constraints are critical for selecting high-quality pseudo-labels and improving hand pose estimation performance.

Amount of Pseudo-label. We analyze the effect of using different fractions of pseudo-labels in semi-supervised learning on the object ADD-0.1D performance on HO-3D dataset [16]. We uniformly sample 20%, 40%, 60%, and 80% fraction of the collected pseudo-labels for semi-supervised learning. As shown in Figure 9, the more pseudo hand labels used in training, the better object performance the model could achieve. Even though the amount of annotated object pose keeps the same in training, the object performance could also be improved for two reasons: First, the better hand performance can help the object pose estimation via contextual reasoning explicitly; Second, the shared encoder is strengthened after semi-supervised learning and boosts object pose implicitly.

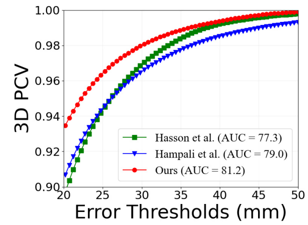


Figure 8: The hand mesh 3D PCV performance on the HO-3D dataset. Our method outperforms previous work [19, 16] by a large margin.

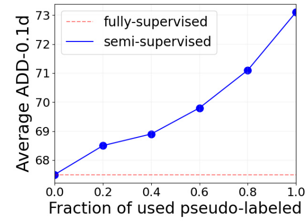


Figure 9: Ablation study of different fractions of pseudo labels contributes to object pose performance gain on the HO-3D dataset.

6.6. Cross-domain Generalization Ability

We evaluate the generalization performance of our model on FPHA [11] and Freihand [76] dataset before and after semi-supervised learning. Follow the protocol of [73, 76], we report the joint AUC and mesh AUC after Procrustes alignment and F-scores [29]. To evaluate the hand mesh performance on the FPHA dataset, we fit the MANO model to the provided ground-truth hand joints following [18]. The generalization results are shown in Table 5. Our model trained with semi-supervised learning has a much more accurate hand joints and mesh estimation compared with the baseline that only trained with supervised learning. In semi-supervised learning, we utilize more training data from the Something-Something video dataset [14] that covers diverse hand poses interacting with objects and subjects in the wild. The model could thus benefit from those data sources in the semi-supervised training stage and yield much better results on generalization across different out-of-domain datasets.

7. Conclusion

In this work, we propose a semi-supervised learning framework for estimating the 3D hand pose and 6-Dof object pose simultaneously, where the hand-object interaction regions are taken as the context for reasoning the object pose. After training the model on the annotated dataset, we deploy it on a large-scale video dataset to generate pseudo hand labels, and then perform spatial-temporal filtering to obtain high-quality ones. Finally, the model is retrained on the union set of real- and pseudo-labels under semi-supervised learning. Experimental results show that our method substantially improves the hand and object pose performance as well as has better cross-domain generalization.

Acknowledgements. This work was supported, in part, by grants from DARPA LwLL, NSF 1730158 CI-New: Cognitive Hardware and Software Ecosystem Community Infrastructure (CHASE-CI), NSF ACI-1541349 CC*DNI Pacific Research Platform, and gifts from Qualcomm and TuSimple.

References

[1] Eric Arazo, Diego Ortego, Paul Albert, Noel E O'Connor, and Kevin McGuinness. Pseudo-labeling and confirmation bias in deep semi-supervised learning. *arXiv preprint arXiv:1908.02983*, 2019.

[2] Seungryul Baek, Kwang In Kim, and Tae-Kyun Kim. Pushing the envelope for rgb-based dense 3d hand pose estimation via neural rendering. In *CVPR*, pages 1067–1076, 2019.

[3] Peter Battaglia, Razvan Pascanu, Matthew Lai, Danilo Jimenez Rezende, et al. Interaction networks for learning about objects, relations and physics. In *Advances in Neural Information Processing Systems*, pages 4502–4510, 2016.

[4] Adnane Boukhayma, Rodrigo de Bem, and Philip HS Torr. 3d hand shape and pose from images in the wild. In *CVPR*, pages 10843–10852, 2019.

[5] Yujun Cai, Liuhao Ge, Jianfei Cai, and Junsong Yuan. Weakly-supervised 3d hand pose estimation from monocular rgb images. In *ECCV*, pages 666–682, 2018.

[6] Xinlei Chen, Abhinav Shrivastava, and Abhinav Gupta. Neil: Extracting visual knowledge from web data. In *ICCV*, pages 1409–1416, 2013.

[7] Y. Chen, Z. Tu, Di Kang, Ruizhi Chen, Linchao Bao, Zhengyou Zhang, and J. Yuan. Joint hand-object 3d reconstruction from a single image with cross-branch feature fusion. *ArXiv*, abs/2006.15561, 2020.

[8] Chiho Choi, Sang Ho Yoon, Chin-Ning Chen, and Karthik Ramani. Robust hand pose estimation during the interaction with an unknown object. In *ICCV*, pages 3123–3132, 2017.

[9] Bardia Doosti, Shujon Naha, M. Mirbagheri, and David J. Crandall. Hope-net: A graph-based model for hand-object pose estimation. *CVPR*, pages 6607–6616, 2020.

[10] Guillermo Garcia-Hernando, Edward Johns, and Tae-Kyun Kim. Physics-based dexterous manipulations with estimated hand poses and residual reinforcement learning. *arXiv preprint arXiv:2008.03285*, 2020.

[11] Guillermo Garcia-Hernando, Shanxin Yuan, Seungryul Baek, and Tae-Kyun Kim. First-person hand action benchmark with rgb-d videos and 3d hand pose annotations. *CVPR*, pages 409–419, 2018.

[12] Liuhao Ge, Zhou Ren, Yuncheng Li, Zehao Xue, Yingying Wang, Jianfei Cai, and Junsong Yuan. 3d hand shape and pose estimation from a single rgb image. In *CVPR*, pages 10833–10842, 2019.

[13] Oliver Glauser, Shihao Wu, Daniele Panozzo, Otmar Hilliges, and Olga Sorkine-Hornung. Interactive hand pose estimation using a stretch-sensing soft glove. *TOG*, 38(4):1–15, 2019.

[14] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fründ, Peter Yianilos, Moritz Mueller-Freitag, Florian Hoppe, Christian Thurau, Ingo Bax, and Roland Memisevic. The "something something" video database for learning and evaluating visual common sense. *ICCV*, pages 5843–5851, 2017.

[15] Klaus Greff, R. Srivastava, J. Koutník, Bastiaan Steunebrink, and J. Schmidhuber. Lstm: A search space odyssey. *IEEE Transactions on Neural Networks and Learning Systems*, 28:2222–2232, 2017.

[16] Shreyas Hampali, Mahdi Rad, Markus Oberweger, and Vincent Lepetit. Honnotate: A method for 3d annotation of hand and objects poses. *CVPR*, 2019.

[17] Ankur Handa, Karl Van Wyk, Wei Yang, Jacky Liang, Yu-Wei Chao, Qian Wan, Stan Birchfield, Nathan Ratliff, and Dieter Fox. Dexpilot: Vision-based teleoperation of dexterous robotic hand-arm system. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 9164–9170. IEEE, 2020.

[18] Yana Hasson, Bugra Tekin, Federica Bogo, Ivan Laptev, Marc Pollefeys, and Cordelia Schmid. Leveraging photometric consistency over time for sparsely supervised hand-object reconstruction. *CVPR*, 2020.

[19] Yana Hasson, Gul Varol, Dimitrios Tzionas, Igor Kalevatykh, Michael J Black, Ivan Laptev, and Cordelia Schmid. Learning joint reconstruction of hands and manipulated objects. In *CVPR*, pages 11807–11816, 2019.

[20] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *CVPR*, pages 2961–2969, 2017.

[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CVPR*, pages 770–778, 2016.

[22] Geoffrey E. Hinton, Oriol Vinyals, and J. Dean. Distilling the knowledge in a neural network. *ArXiv*, abs/1503.02531, 2015.

[23] Han Hu, Jiayuan Gu, Zheng Zhang, Jifeng Dai, and Yichen Wei. Relation networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3588–3597, 2018.

[24] Yinlin Hu, Joachim Hugonot, Pascal Fua, and Mathieu Salzmann. Segmentation-driven 6d object pose estimation. *CVPR*, pages 3380–3389, 2019.

[25] Wolfgang Hürst and Casper Van Wezel. Gesture-based interaction via finger tracking for mobile augmented reality. *Multimedia Tools and Applications*, 62(1):233–258, 2013.

[26] Umar Iqbal, Pavlo Molchanov, Thomas Breuel Juergen Gall, and Jan Kautz. Hand pose estimation via latent 2.5 d heatmap regression. In *ECCV*, pages 118–134, 2018.

[27] Ahmet Iscen, Giorgos Tolias, Yannis Avrithis, and Ondrej Chum. Label propagation for deep semi-supervised learning. In *CVPR*, pages 5070–5079, 2019.

[28] Wadim Kehl, Fabian Manhardt, Federico Tombari, Slobodan Ilic, and Nassir Navab. Ssd-6d: Making rgb-based 3d detection and 6d pose estimation great again. *ICCV*, pages 1530–1538, 2017.

[29] Arno Knapitsch, Jaesik Park, Qian-Yi Zhou, and Vladlen Koltun. Tanks and temples: Benchmarking large-scale scene reconstruction. *ACM Transactions on Graphics (ToG)*, 36(4):1–13, 2017.

[30] Dominik Kulon, Riza Alp Guler, Iasonas Kokkinos, Michael M. Bronstein, and Stefanos Zafeiriou. Weakly-supervised mesh-convolutional hand reconstruction in the wild. In *CVPR*, June 2020.

[31] Dong-Hyun Lee. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks.

In *Workshop on challenges in representation learning, ICML*, volume 3, page 2, 2013.

[32] Xiaolong Li, He Wang, Li Yi, Leonidas J Guibas, A Lynn Abbott, and Shuran Song. Category-level articulated object pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3706–3715, 2020.

[33] Tsung-Yi Lin, Piotr Dollár, Ross B. Girshick, Kaiming He, Bharath Hariharan, and Serge J. Belongie. Feature pyramid networks for object detection. *CVPR*, pages 936–944, 2017.

[34] Shaowei Liu, Guijin Wang, Pengwei Xie, and Cairong Zhang. Light and fast hand pose estimation from spatial-decomposed latent heatmap. *IEEE Access*, 8:53072–53081, 2020.

[35] David Lopez-Paz, Léon Bottou, Bernhard Schölkopf, and Vladimir Vapnik. Unifying distillation and privileged information. *arXiv preprint arXiv:1511.03643*, 2015.

[36] Gyeongsik Moon, Ju Yong Chang, and Kyoung Mu Lee. V2v-posenet: Voxel-to-voxel prediction network for accurate 3d hand and human pose estimation from a single depth map. In *Proceedings of the IEEE conference on computer vision and pattern Recognition*, pages 5079–5088, 2018.

[37] Franziska Mueller, Florian Bernard, Oleksandr Sotnychenko, Dushyant Mehta, Srinath Sridhar, Dan Casas, and Christian Theobalt. Ganerated hands for real-time 3d hand tracking from monocular rgb. In *CVPR*, page 11, 2018.

[38] Franziska Mueller, Dushyant Mehta, Oleksandr Sotnychenko, Srinath Sridhar, Dan Casas, and Christian Theobalt. Real-time hand tracking under occlusion from an egocentric rgb-d sensor. In *ICCVW*, pages 1284–1293, 2017.

[39] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *ECCV*, pages 483–499. Springer, 2016.

[40] Markus Oberweger, Paul Wohlhart, and Vincent Lepetit. Generalized feedback loop for joint hand-object pose estimation. *IEEE TPAMI*, 2019.

[41] Iason Oikonomidis, Nikolaos Kyriazis, and Antonis A Argyros. Full dof tracking of a hand interacting with an object by modeling occlusions and physical constraints. In *ICCV*, pages 2088–2095. IEEE, 2011.

[42] Georgios Pavlakos, Luyang Zhu, Xiaowei Zhou, and Kostas Daniilidis. Learning to estimate 3d human pose and shape from a single color image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 459–468, 2018.

[43] Sida Peng, Yuan Liu, Qi-Xing Huang, Hujun Bao, and Xiaowei Zhou. Pvnet: Pixel-wise voting network for 6dof pose estimation. *CVPR*, pages 4556–4565, 2019.

[44] Thammathip Piumsomboon, Adrian Clark, Mark Billinghurst, and Andy Cockburn. User-defined gestures for augmented reality. In *IFIP Conference on Human-Computer Interaction*, pages 282–299, 2013.

[45] 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. *ICCV*, pages 3848–3856, 2017.

[46] Ilija Radosavovic, Piotr Dollár, Ross Girshick, Georgia Gkioxari, and Kaiming He. Data distillation: Towards omni-supervised learning. In *CVPR*, pages 4119–4128, 2018.

[47] Joseph Redmon, Santosh Kumar Divvala, Ross B. Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. *CVPR*, pages 779–788, 2016.

[48] Ellen Riloff and Janyce Wiebe. Learning extraction patterns for subjective expressions. In *EMNLP*, pages 105–112, 2003.

[49] Grégory Rogez, Maryam Khademi, JS Supančič III, Jose Maria Martinez Montiel, and Deva Ramanan. 3d hand pose detection in egocentric rgb-d images. In *ECCV*, pages 356–371, 2014.

[50] Javier Romero, Dimitrios Tzionas, and Michael J Black. Embodied hands: Modeling and capturing hands and bodies together. *ToG*, 36(6):245, 2017.

[51] Chuck Rosenberg, Martial Hebert, and Henry Schneiderman. Semi-supervised self-training of object detection models. *WACV/MOTION*, 2, 2005.

[52] Amir Sadeghian, Alexandre Alahi, and Silvio Savarese. Tracking the untrackable: Learning to track multiple cues with long-term dependencies. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 300–311, 2017.

[53] Adam Santoro, David Raposo, David G Barrett, Mateusz Malinowski, Razvan Pascanu, Peter Battaglia, and Timothy Lillicrap. A simple neural network module for relational reasoning. volume 30, 2017.

[54] Max Schwarz, Hannes Schulz, and Sven Behnke. Rgb-d object recognition and pose estimation based on pre-trained convolutional neural network features. In *2015 IEEE international conference on robotics and automation (ICRA)*, pages 1329–1335. IEEE, 2015.

[55] H Scudder. Probability of error of some adaptive pattern-recognition machines. *IEEE Transactions on Information Theory*, 11(3):363–371, 1965.

[56] Weiwei Shi, Yihong Gong, Chris Ding, Zhiheng MaXiaoyu Tao, and Nanning Zheng. Transductive semi-supervised deep learning using min-max features. In *ECCV*, pages 299–315, 2018.

[57] Adrian Spurr, Jie Song, Seonwook Park, and Otmar Hilliges. Cross-modal deep variational hand pose estimation. In *CVPR*, 2018.

[58] Srinath Sridhar, Anna Maria Feit, Christian Theobalt, and Antti Oulasvirta. Investigating the dexterity of multi-finger input for mid-air text entry. In *CHI*, pages 3643–3652, 2015.

[59] Bugra Tekin, Federica Bogo, and Marc Pollefeys. H+o: Unified egocentric recognition of 3d hand-object poses and interactions. In *CVPR*, pages 4511–4520, 2019.

[60] Bugra Tekin, Sudipta N. Sinha, and Pascal Fua. Real-time seamless single shot 6d object pose prediction. *CVPR*, pages 292–301, 2018.

[61] Xiaolong Wang, Ross B. Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. *CVPR*, pages 7794–7803, 2018.

[62] Xiaolong Wang and Abhinav Gupta. Videos as space-time region graphs. In *Proceedings of the European conference on computer vision (ECCV)*, pages 399–417, 2018.

[63] Nicholas Watters, Daniel Zoran, Theophane Weber, Peter Battaglia, Razvan Pascanu, and Andrea Tacchetti. Visual interaction networks: Learning a physics simulator from video. *Advances in neural information processing systems*, 30:4539–4547, 2017.

[64] Chao-Yuan Wu, Christoph Feichtenhofer, Haoqi Fan, Kaiming He, Philipp Krahenbuhl, and Ross Girshick. Long-term feature banks for detailed video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 284–293, 2019.

[65] Sitao Xiang and Hao Li. Revisiting the continuity of rotation representations in neural networks. *ArXiv*, abs/2006.06234, 2020.

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

[67] Yu Xiang, Tanner Schmidt, Venkatraman Narayanan, and Dieter Fox. Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes. *ArXiv*, abs/1711.00199, 2018.

[68] Qizhe Xie, Eduard Hovy, Minh-Thang Luong, and Quoc V Le. Self-training with noisy student improves imagenet classification. *arXiv preprint arXiv:1911.04252*, 2019.

[69] Chi Xu, Lakshmi Narasimhan Govindarajan, Yu Zhang, and Li Cheng. Lie-x: Depth image based articulated object pose estimation, tracking, and action recognition on lie groups. *International Journal of Computer Vision*, 123:454–478, 2017.

[70] Yuan Yao and Yun Fu. Real-time hand pose estimation from rgb-d sensor. In *ICME*, pages 705–710, 2012.

[71] David Yarowsky. Unsupervised word sense disambiguation rivaling supervised methods. In *ACL*, pages 189–196, 1995.

[72] Shanxin Yuan, Guillermo Garcia-Hernando, Björn Stenger, Gyeongsik Moon, Ju Yong Chang, Kyoung Mu Lee, Pavlo Molchanov, Jan Kautz, Sina Honari, Liuhao Ge, et al. Depth-based 3d hand pose estimation: From current achievements to future goals. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2636–2645, 2018.

[73] Xiong Zhang, Qiang Li, Hong Mo, Wenbo Zhang, and Wen Zheng. End-to-end hand mesh recovery from a monocular rgb image. In *CVPR*, pages 2354–2364, 2019.

[74] Yuxiao Zhou, Marc Habermann, Weipeng Xu, Ikhsanul Habibie, Christian Theobalt, and Feng Xu. Monocular real-time hand shape and motion capture using multi-modal data. *CVPR*, 2020.

[75] Christian Zimmermann and Thomas Brox. Learning to estimate 3d hand pose from single rgb images. In *CVPR*, pages 4903–4911, 2017.

[76] Christian Zimmermann, Duygu Ceylan, Jimei Yang, Bryan Russell, Max Argus, and Thomas Brox. Freihand: A dataset for markerless capture of hand pose and shape from single rgb images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 813–822, 2019.