

S²Contact: Graph-based Network for 3D Hand-Object Contact Estimation with Semi-Supervised Learning

Tze Ho Elden Tse^{1*}, Zhongqun Zhang^{1*}, Kwang In Kim², Aleš Leonardis¹,
Feng Zheng³, and Hyung Jin Chang¹

¹ University of Birmingham, UK

² UNIST, Korea

³ SUSTech, China

Abstract. Despite the recent efforts in accurate 3D annotations in hand and object datasets, there still exist gaps in 3D hand and object reconstructions. Existing works leverage contact maps to refine inaccurate hand-object pose estimations and generate grasps given object models. However, they require explicit 3D supervision which is seldom available and therefore, are limited to constrained settings, e.g., where thermal cameras observe residual heat left on manipulated objects. In this paper, we propose a novel semi-supervised framework that allows us to learn contact from monocular images. Specifically, we leverage visual and geometric consistency constraints in large-scale datasets for generating pseudo-labels in semi-supervised learning and propose an efficient graph-based network to infer contact. Our semi-supervised learning framework achieves a favourable improvement over the existing supervised learning methods trained on data with ‘limited’ annotations. Notably, our proposed model is able to achieve superior results with less than half the network parameters and memory access cost when compared with the commonly-used PointNet-based approach. We show benefits from using a contact map that rules hand-object interactions to produce more accurate reconstructions. We further demonstrate that training with pseudo-labels can extend contact map estimations to out-of-domain objects and generalise better across multiple datasets. Project page is available.¹

1 Introduction

Understanding hand-object interactions have been an active area of study in recent years [18, 16, 17, 3, 49, 26, 20, 62, 31, 51]. Besides common practical applications in augmented and virtual reality [15, 54, 36, 51, 50, 66], it is a key ingredient to advanced human-computer interaction [52] and imitation learning in robotics [64]. In this paper, as illustrated in Figure 1, we tackle the problem of 3D reconstruction of the hand and manipulated object with the focus on contact map estimation.

* Equal contribution

¹ <https://eldentse.github.io/s2contact/>

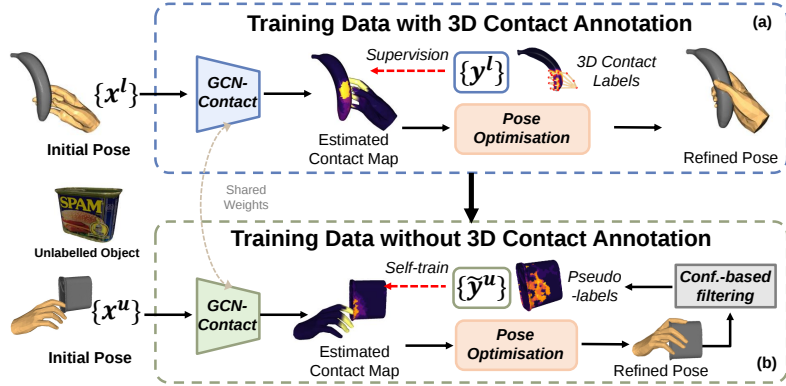


Fig. 1. Overview of our semi-supervised learning framework, S²Contact. (a) The model is pre-trained on a small annotated dataset. (b) Then, it is deployed on unlabelled datasets to collect pseudo-labels. The pseudo-labels are filtered with confidence-based on visual and geometric consistencies. Upon predicting the contact map, the hand and object poses are jointly optimised to achieve target contact via a contact model [12].

Previous works in hand-object interactions typically formulate this as a joint hand and object pose estimation problem. Along with the development of data collection and annotation methods, more accurate 3D annotations for real datasets [11, 14, 4] are available for learning-based methods [10, 49]. Despite the efforts, there still exist gaps between hand-object pose estimation and contact as ground-truth in datasets are not perfect. Recent works attempt to address this problem with interaction constraints (attraction and repulsion) under an optimisation framework [17, 3, 62]. However, inferred poses continue to exhibit sufficient error to cause unrealistic hand-object contact, making downstream tasks challenging [12]. In addition, annotations under constrained laboratory environments rely on strong priors such as limited hand motion which prevents the trained model from generalising to novel scenes and out-of-domain objects.

To address the problem of hand-object contact modelling, Brahmhatt *et al.* [1] used thermal cameras observing the heat transfer from hand to object after the grasp to capture detailed ground-truth contact. Their follow-up work contributed a large grasp dataset (*ContactPose*) with contact maps and hand-object pose annotations. Recent works are able to leverage contact maps to refine inaccurate hand-object pose estimations [12] and generate grasps given object model [20]. Therefore, the ability to generate an accurate contact map is one of the key elements to reasoning physical contact. However, the number of annotated objects is incomparable to manipulated objects in real life and insufficient to cover a wide range of human intents. Furthermore, obtaining annotations for contact maps is non-trivial as it requires thermal sensors during data collection.

To enable the wider adoption of contact maps, we propose a unified framework that leverages existing hand-object datasets for generating pseudo-labels in semi-supervised learning. Specifically, we propose to exploit the visual and ge-

ometric consistencies of contact maps in hand-object interactions. This is built upon the idea that the poses of the hands and objects are highly-correlated where the 3D pose of the hand often indicates the orientation of the manipulated object. We further extend this by enforcing our contact consistency loss for the contact maps across a video.

As the input to contact map estimator are in the form of point clouds, recent related works [12, 20] typically follow a PointNet-based architectures [38, 39]. This achieves permutation invariance of points by operating on each point independently and subsequently applying a symmetric function to accumulate features [55]. However, the network performances are limited as points are treated independently at a local scale to maintain permutation invariance. To overcome this fundamental limitation, many recent approaches adopt graph convolutional networks (GCN) [9, 25] and achieve state-of-the-art performances in 3D representation learning on point clouds for classification, part segmentation and semantic segmentation [55, 28, 30]. The ability to capture local geometric structures while maintaining permutation invariance is particularly important for estimating contact maps. However, it comes at the cost of high computation and memory usage for constructing a local neighbourhood with K -nearest neighbour (K -NN) search on point clouds at each training epoch. For this reason, we design a graph-based neural network that demonstrates superior results with less than half the learning parameters and faster convergence.

Our contributions are three-fold:

- We propose a novel semi-supervised learning framework that combines pseudo-label with consistency training. Experimental results demonstrate the effectiveness of this training strategy.
- We propose a novel graph-based network for processing hand-object point clouds, which is at least two times more efficient than PointNet-based architecture for estimating contact between hand and object.
- We conduct comprehensive experiments on three commonly-used hand-object datasets. Experiments show that our proposed framework S²Contact outperforms recent semi-supervised methods.

2 Related work

Our work tackles the problem of hand and object reconstruction from monocular RGB videos, exploiting geometric and visual consistencies on contact maps for semi-supervised learning. To the best of our knowledge, we are the first to apply such consistencies on hand-object scenarios. We first review the literature on *hand-object reconstruction*. Then, we review *point cloud analysis* with the focus of graph-based methods. Finally, we provide a brief review on *semi-supervised learning in 3D hand-object pose estimation*.

2.1 Hand-object reconstruction

Previous works mainly tackle 3D pose estimations on hands [42, 67, 35, 44, 60, 60, 47] and objects [27, 29, 58, 5, 6] separately. Joint reconstruction of hands and

objects has been receiving increasing attention [18, 16, 3, 17]. Hasson *et al.* [18] introduces an end-to-end model to regress MANO hand parameters jointly with object mesh vertices deformed from a sphere and incorporates contact losses which encourages contact surfaces and penalises penetrations between hand and object. A line of works [49, 10, 16, 3, 17, 62, 12, 19] assume known object models and regress a 6DoF object pose instead. Other works focus on grasp synthesis [8, 21, 46, 20]. In contrast, our method is in line with recent optimisation-based approaches for modelling 3D hand-object contact. ContactOpt [12] proposes a contact map estimation network and a contact model to produce realistic hand-object interaction. ContactPose dataset [2] is unique in capturing ground-truth thermal contact maps. However, 3D contact labels are seldom available and limited to constrained laboratory settings. In this work, we treat contact maps as our primary learning target and leverage unannotated datasets.

2.2 Point cloud analysis

Since point cloud data is irregular and unordered, early works tend to project the original point clouds to intermediate voxels [33] or images [63], *i.e.* translating into a well-explored 2D image problem. As information loss caused by projection degrades the representational quality, PointNet [38] is proposed to directly process unordered point sets and PointNet++ [39] extends on local point representation in multi-scale. As PointNet++ [39] can be viewed as the generic point cloud analysis network framework, the research focus has been shifted to generating better regional points representation. Methods can be divided into convolution [57, 59], graph [55, 28, 30] and attention [13, 65]-based. **Graph-based methods.** GCNs have been gaining much attention in the last few years. This is due to two reasons: 1) the rapid increase of non-Euclidean data in real-world applications and 2) the limited performance of convolutional neural networks when dealing with such data. As the unstructured nature of point clouds poses a representational challenge in the community, graph-based methods treat points as nodes of a graph and formulate edges according to their spatial/feature relationships. MoNet [34] defines the convolution as Gaussian mixture models in a local pseudo-coordinate system. 3D-GCN [30] proposes a deformable kernels which has shift and scale-invariant properties for point cloud processing. DGCNN [55] proposes to gather nearest neighbouring points in feature space and follow by the EdgeConv operators for feature extraction. The EdgeConv operator dynamically computes node adjacency at each graph layer using the distance between point features. In this paper, we propose a computationally efficient network for contact map estimation which requires less than half the parameters of PointNet [38] and GPU memory of DGCNN [55].

2.3 Semi-supervised learning in 3D hand-object pose estimation

Learning from both labelled and unlabelled data simultaneously has recently attracted growing interest in 3D hand pose estimation [61, 43, 23, 7, 48]. They typically focus on training models with a small amount of labelled data as well as a

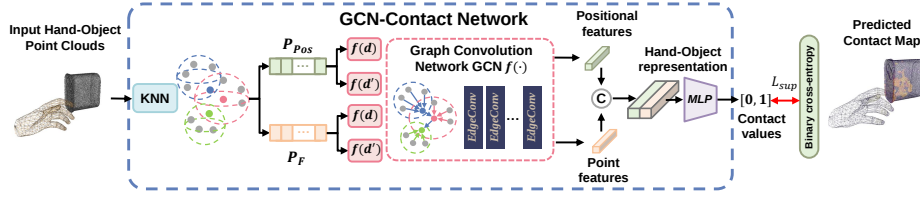


Fig. 2. Framework of GCN-Contact. The network takes hand-object point clouds $\mathbf{P} = (\mathbf{P}_{hand}, \mathbf{P}_{obj})$ as input and perform K -NN search separately on 3D position $\mathbf{P}_{pos}$ and point features $\mathbf{P}_F$, *i.e.* $\mathbf{P} = \{\mathbf{P}_{pos}, \mathbf{P}_F\}$. Different dilation factors d, d' are used to enlarge the receptive field for graph convolution $f(\cdot)$. Finally, features are concatenated and pass to MLP to predict contact map.

relatively larger amount of unlabelled data. After training on human-annotated datasets, pseudo-labelling and consistency training can be used to train further and a teacher-student network with exponential moving average (EMA) strategy [53] is common to accelerate the training. For instance, So-HandNet [7] leverages the consistency between the recovered hand point cloud and the original hand point cloud for semi-supervised training. SemiHand [61] is the first to combine pseudo-labelling and consistency learning for hand pose estimation. Liu *et al.* [31] is the only prior work on 3D hand-object pose estimation with semi-supervised learning. They proposed spatial and temporal constraints for selecting the pseudo-labels from videos. However, they are limited to pseudo hand labels and did not account for physical contact with manipulated objects. In contrast, our work is the first to explore pseudo-labelling for 3D hand-object contact map with geometric and visual consistency constraints.

3 Methodology

Given a noisy estimate of hand and object meshes from an image-based algorithm, we seek to learn a hand-object contact region estimator by exploiting real-world hand and object video datasets without contact ground-truths. Figure 1 shows an overview of our approach. In the following section, we describe our learned contact map estimation network (GCN-Contact) in Section 3.1 and our newly proposing semi-supervised training pipeline (S²Contact) in Section 3.2 that utilise a teacher-student mutual learning framework.

3.1 GCN-Contact: 3D hand-object contact estimation

As pose estimates from an image-based algorithm can be potentially inaccurate, GCN-Contact learns to infer contact maps $\mathbf{C} = (\mathbf{C}_{hand}, \mathbf{C}_{obj})$ from hand and object point clouds $\mathbf{P} = (\mathbf{P}_{hand}, \mathbf{P}_{obj})$. We adopted the differential MANO [41] model from [18]. It maps pose ($\boldsymbol{\theta} \in \mathbb{R}^{51}$) and shape ($\boldsymbol{\beta} \in \mathbb{R}^{10}$) parameters to a mesh with $N = 778$ vertices. Pose parameters ($\boldsymbol{\theta}$) consists of 45 DoF (*i.e.* 3 DoF for each of the 15 finger joints) plus 6 DoF for rotation and translation of the

wrist joint. Shape parameters (β) are fixed for a given person. We sample 2048 points randomly from object model to form object point cloud. Following [12], we include F -dimensional point features for each point: binary per-point feature indicating whether the point belongs to the hand or object, distances from hand to object and surface normal information. With network input $\mathbf{P} = (\mathbf{P}_{hand}, \mathbf{P}_{obj})$ where $\mathbf{P}_{hand} \in \mathbb{R}^{778 \times F}$ and $\mathbf{P}_{obj} \in \mathbb{R}^{2048 \times F}$, GCN-Contact can be trained to infer discrete contact representation ($\mathbf{C} = (\mathbf{C}_{hand}, \mathbf{C}_{obj}) \in [0, 1]$) [2] using binary cross-entropy loss. Similarly to [12], the contact value range $[0, 1]$ is evenly split into 10 bins and the training loss is weighted to account for class imbalance.

Revisiting PointNet-based methods. Recent contact map estimators are based on PointNet [20] and PointNet++ [12]. PointNet [38] directly processes unordered point sets using shared multi-layer perceptron (MLP) networks. PointNet++ [39] learns hierarchical features by stacking multiple learning stages and recursively capturing local geometric structures. At each learning stage, farthest point sampling (FPS) algorithm is used to re-sample a fixed number of points and K neighbours are obtained from ball query’s local neighbourhood for each sampled point to capture local structures. The kernel operation of PointNet++ for point $p_i \in \mathbb{R}^F$ with F -dimensional features can be described as:

$$\hat{p}_i = \sigma(\Phi(p_j | j \in \mathcal{N}(i))), \quad (1)$$

where the updated point $\hat{p}_i$ is formed by max-pooling function $\sigma(\cdot)$ and PointNet as the basic building block for local feature extractor $\Phi(\cdot)$ around point neighbourhood $\mathcal{N}(i)$ of point p_i . The kernel of the point convolution can be implemented with MLPs. However, MLPs are unnecessarily performed on the neighbourhood features which causes a considerable amount of latency in PointNet++ [40]. This motivates us to employ advanced local feature extractors such as convolution [57, 59], graph [55, 28, 30] or self-attention mechanisms [13, 65].

Local geometric information. While contact map estimation can take advantage of detailed local geometric information, they usually suffer from two major limitations. First, the computational complexity is largely increased with delicate extractors which leads to low inference latency. For instance, in graph-based methods, neighbourhood information gathering modules are placed for better modelling of the locality on point clouds. This is commonly established by K -nearest neighbour (K -NN) search which increases the computational cost quadratically with the number of points and even further for dynamic feature representation [55]. For reference on ModelNet40 point cloud classification task [40], the inference speed of PointNet [38] is 41 times faster than DGCNN [55]. Second, Liu *et al.*’s investigation on local aggregation operators reveals that advanced local feature extractors make surprisingly similar contributions to the network performance under the same network input [32]. For these reasons, we are encouraged to develop a computationally efficient design while maintaining comparable accuracy for learning contact map estimation.

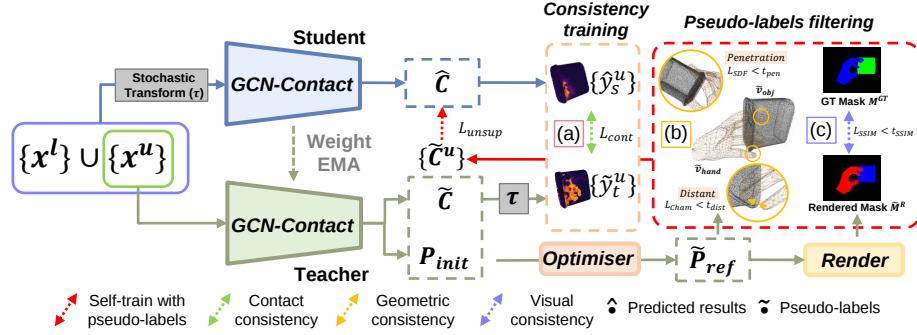


Fig. 3. S²Contact pipeline. We adopt our proposed graph-based network GCN-Contact as backbone. We utilise a teacher-student mutual learning framework which is composed of a learnable student and an EMA teacher. The student network is trained with labelled data $\{x^l, y^l\}$. For unlabelled data x^u , the student network takes pseudo contact labels $\tilde{C}^u$ from its EMA teacher and compares with its predictions $\hat{C}$. Please note that pseudo contact labels $\tilde{C}^u$ is a subset of $\hat{C}$, i.e. $\tilde{C}^u \in \hat{C}$. (a) refers to contact consistency constraint for consistency training. To improve the quality of pseudo-label, we adopt a confidence-based filtering mechanism to geometrically (b) and visually (c) filter out predictions that violate contact constraints.

Proposed method. To overcome the aforementioned limitations, we present a simple yet effective graph-based network for contact map estimation. We use EdgeConv [55] to generate edge features that describe the relationships between a point and its neighbours:

$$\Phi(p_i, p_j) = \text{ReLU}(\text{MLP}(p_j - p_i, p_i)), \quad j \in \mathcal{N}(i), \quad (2)$$

where neighbourhood $\mathcal{N}(i)$ is obtained by K -NN search around the point p_i . As shown in Figure 2, we only compute K -NN search once at each network pass to improve computational complexity and reduce memory usage. In addition, we apply dilation on the K -NN results to increase the receptive field without loss of resolution. To better construct local regions when hand and object are perturbed, we propose to perform K -NN search on 3D position and point features separately. Note that [12] perform ball query on $0.1 - 0.2m$ radius and [55] combine both position and features. Finally, we take inspirations from the Inception model [45] in which they extract multi-scale information by using different kernel sizes in different paths of the architecture. Similarly, we process spatial information at various dilation factors and then aggregates. The experiment demonstrates the effectiveness of our proposed method and is able to achieve constant memory access cost regardless of the size of dilation factor d (see Table 1).

3.2 S²Contact: Semi-supervised training pipeline

Collecting ground-truth contact annotation for hand-object dataset can be both challenging and time-consuming. To alleviate this, we introduce a semi-supervised

learning framework to learn 3D hand-object contact estimation by leveraging large-scale unlabelled videos. As shown in Figure 1, our proposed framework relies on two training stages: 1) pre-training stage where the model is pre-trained on the existing labelled data [2]; 2) semi-supervised stage where the model is trained by the pseudo-labels from unlabelled hand-object datasets [4, 14, 11]. As pseudo-labels are often noisy, we propose confidence-based filtering with geometric and visual consistency constraints to improve the quality of pseudo-labels.

Pre-training. As good initial contact estimate enables semi-supervision, we pre-train our graph-based contact estimator using a small labelled dataset $\{\mathbf{x}^l, \mathbf{y}^l\}$. We followed [12] and optimise hand-object poses to achieve target contact. Upon convergence, we clone the network to create a pair of student-teacher networks.

Pseudo-label generation. To maintain a reliable performance margin over the student network throughout the training, we adopt an EMA teacher which is commonly used in semi-supervised learning. The output of the student network is the predicted contact map $\hat{\mathbf{C}}$. The teacher network generates pseudo-labels which includes pre-filter contact map $\tilde{\mathbf{C}}$ and refined hand-object pose $\tilde{P}_{ref}$. As it is crucial for the teacher network to generate high-quality pseudo-labels under a semi-supervised framework, we propose a confidence-based filtering mechanism that leverages geometric and visual consistency constraints.

Contact consistency constraint for consistency training. We propose a contact consistency loss to encourage robust and stable predictions for unlabelled data $\mathbf{x}^u$. As shown in Figure 3 (a), we first apply stochastic transformations $\mathcal{T}$ which includes flipping, rotation and scaling on the input hand-object point clouds $\mathbf{x}^u$ for the student network. The predictions of the student network $\hat{\mathbf{y}}_s^u \in \hat{\mathbf{C}}$ are compared with the teacher predictions $\tilde{\mathbf{y}}_t^u \in \tilde{\mathbf{C}}$ processed by the same transformation $\mathcal{T}$ using contact consistency loss:

$$\begin{aligned} \mathcal{L}_{cont}(\mathbf{x}^u) &= \|\Omega(\mathcal{T}(\mathbf{x}^u)) - (\mathcal{T}(\Omega(\mathbf{x}^u)))\|_1 \\ &= \|\hat{\mathbf{y}}_s^u - \tilde{\mathbf{y}}_t^u\|_1, \end{aligned} \quad (3)$$

where $\Omega(\cdot)$ represents the predicted contact map.

Geometric consistency constraint for pseudo-labels filtering. As shown in Figure 3 (b), we propose a geometric consistency constraint to the hand and object pseudo pose label $\tilde{P}_{ref}$. Concretely, we allow the Chamfer distance $\mathcal{L}_{Cham}$ between hand and object meshes to be less than threshold t_{dist} :

$$\mathcal{L}_{Cham}(\tilde{\mathbf{v}}_{hand}, \tilde{\mathbf{v}}_{obj}) = \frac{1}{|\tilde{\mathbf{v}}_{obj}|} \sum_{x \in \tilde{\mathbf{v}}_{obj}} d_{\tilde{\mathbf{v}}_{hand}}(x) + \frac{1}{|\tilde{\mathbf{v}}_{hand}|} \sum_{y \in \tilde{\mathbf{v}}_{hand}} d_{\tilde{\mathbf{v}}_{obj}}(y), \quad (4)$$

where $\tilde{\mathbf{v}}_{hand}$ and $\tilde{\mathbf{v}}_{obj}$ refers to hand and object point sets, $d_{\tilde{\mathbf{v}}_{hand}}(x) = \min_{y \in \tilde{\mathbf{v}}_{hand}} \|x - y\|_2^2$, and $d_{\tilde{\mathbf{v}}_{obj}}(y) = \min_{x \in \tilde{\mathbf{v}}_{obj}} \|x - y\|_2^2$. Similarly for interpenetration, we use $\mathcal{L}_{SDF}(\tilde{\mathbf{v}}_{obj}) = \sum_{hand, obj} \sum_i \Psi_h(\tilde{\mathbf{v}}_{obj}^i) \leq t_{pen}$ to ensure object is being manipulated by hand. Ψ_h is the Signed Distance Field (SDF) from the hand mesh (i.e., $\Psi_h(\tilde{\mathbf{v}}_{obj}) = -\min(\text{SDF}(\tilde{\mathbf{v}}_{obj}), 0)$) to detect object penetrations.

Visual consistency constraint for pseudo-labels filtering. We observed that geometric consistency is insufficient to correct hand grasp (see Table 5). To address this, we propose a visual consistency constraint to filter out the pseudo-labels whose rendered hand-object image I_{ho} does not match the input image. We first use a renderer [22] to render the hand-object image from the refined pose $\tilde{P}_{ref}$ and obtain the hand-object segment of the input image I by applying the segmentation mask M_{gt} . Then, the structural similarity (SSIM) [56] between two images can be computed. We keep pseudo-labels when $\mathcal{L}_{SSIM} \leq t_{SSIM}$:

$$\mathcal{L}_{SSIM}(I, M_{gt}, \tilde{I}_{ho}) = 1 - SSIM(I \odot M_{gt}, \tilde{I}_{ho}), \quad (5)$$

where $\odot$ denotes element-wise multiplication.

Self-training with pseudo-labels. After filtering pseudo-labels, our model is trained with the union set of the human-annotated dataset and the remaining pseudo-labels. The total loss $\mathcal{L}_{semi}$ can be described as:

$$\mathcal{L}_{semi}(\hat{\mathbf{C}}, \mathbf{y}^l, \tilde{\mathbf{C}}^u, \mathbf{x}^u) = \mathcal{L}_{sup}(\hat{\mathbf{C}}, \mathbf{y}^l) + \mathcal{L}_{unsup}(\hat{\mathbf{C}}, \tilde{\mathbf{C}}^u) + \lambda_c \mathcal{L}_{cont}(\mathbf{x}^u), \quad (6)$$

where $\mathcal{L}_{sup}$ is a supervised contact loss, $\mathcal{L}_{unsup}$ is a unsupervised contact loss with pseudo-labels and λ_c is a hyperparameter. Note that $\mathcal{L}_{sup}$ (see Figure 2) and $\mathcal{L}_{unsup}$ (see Figure 3) are both binary cross-entropy loss.

4 Experiments

Implementation details. We implement our method in PyTorch [37]. All experiments are run on an Intel i9-CPU @ 3.50GHZ, 16 GB RAM, and one NVIDIA RTX 3090 GPU. For pseudo-labels filtering, $t_{dist} = 0.7$, $t_{pen} = 6$ and $t_{SSIM} = 0.25$ are the constant thresholds and stochastic transformations includes flipping ($\pm 20\%$), rotation ($\pm 180^\circ$) and scaling ($\pm 20\%$). We train all parts of the network simultaneously with Adam optimiser [24] at a learning rate 10^{-3} for 100 epochs. We empirically fixed $K = 10$, $d = 4$ to produce the best results.

Datasets and evaluation metrics. *ContactPose* is the first dataset [2] of hand-object contact paired with hand pose, object pose and RGB-D images. It contains 2,306 unique grasps of 25 household objects grasped with 2 functional intents by 50 participants, and more than 2.9M RGB-D grasp images. For fair comparisons with ContactOpt [12], we follow their *Perturbed ContactPose* dataset where hand meshes are modified by additional noise to MANO parameters. This results in 22,624 training and 1,416 testing grasps. *DexYCB* is a recent real dataset for capturing hand grasping of objects [4]. It consists of a total of 582,000 image frames on 20 objects from the YCB-Video dataset [58]. We present results on their default official dataset split settings. *HO-3D* [14] is similar to *DexYCB* where it consists of 78,000 images frames on 10 objects. We present results on the official dataset split (version 2). The hand mesh error is reported after procrustes alignment and in *mm*.

Table 1. Hand error rates (mm) on *Perturbed ContactPose* and *ContactPose* datasets.

	Baseline		DGCNN [55]		Ours	
	joint ↓	mesh ↓	joint ↓	mesh ↓	joint ↓	mesh ↓
<i>Perturbed ContactPose</i>	32.988	33.147	32.592	32.762	29.442	30.635
<i>ContactPose</i>	8.880	8.769	8.767	32.988	5.878	5.765

- *Hand error*: We report the mean end-point error (mm) over 21 joints and mesh error in mm .
- *Object error*: We report the percentage of average object 3D vertices error within 10% of object diameter (ADD-0.1D).
- *Hand-object interaction*: We report the intersection volume (cm^3) and contact coverage (%). Intersection volume is obtained by voxelising the hand and object using a voxel size of $0.5cm$. Contact coverage refers to the percentage of hand points between $\pm 2mm\%$ of the object surface [12].

Baseline. For refining image-based pose estimates, we use the baseline pose estimation network from Hasson *et al.* [16] and retrain it on the training split of the respecting dataset. We filter out frames where the minimum distance between the ground truth hand and object surfaces is greater than $2\ mm$. We also use the contact estimation network DeepContact from Grady *et al.* [12] which takes ground-truth object class and pose. For semi-supervised learning, we use the baseline method from Liu *et al.* [31], a semi-supervised learning pipeline for 3D hand-object pose estimation from large-scale hand-object interaction videos.

4.1 Comparative results

Refining small and large inaccuracies. We use *ContactPose* to evaluate GCN-Contact for refining poses with small (*ContactPose*) and large (*Perturbed ContactPose*) inaccuracies. Table 1 shows the results for both cases. For *Perturbed ContactPose*, the mean end-point error over 21 joints is $82.947mm$ before refinement. This is aimed at testing the ability to improve hand poses with large errors. In contrast, *ContactPose* is used to evaluate mm -scale refinement. As shown, our method consistently outperforms baseline and DGCNN [55]. We attribute the performance gain to multi-scale feature aggregation with dilation. Qualitative comparison with ContactOpt [12] is provided in Figure 4.

Refining Image-based pose estimates. We evaluate S²Contact in refining poses from an image-based pose estimator. We use the baseline image-based pose estimation network from Hasson *et al.* [16] and retrain it on the training split of the respecting dataset. Unlike [12], we do not rely on ground-truth object class and pose. In particular, we compare with the current state-of-the-art [31] which is also a semi-supervised framework for 3D hand-object pose estimation. Liu *et al.* [31] proposes spatial-temporal consistency in large-scale hand-object videos

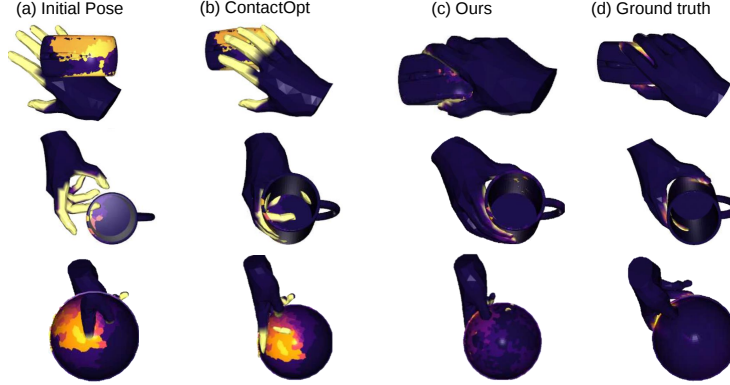


Fig. 4. Qualitative comparison with ContactOpt [12] on *ContactPose*. We observed that penetrations across hand and object (can be seen in (b)) is likely to be caused by contact predictions appearing on both object surfaces. Our model, trained only on *ContactPose*, shows robustness to various hand poses and objects.

Table 2. Error rates on *HO-3D*. Note that Liu *et al.* [31] is the current state-of-the-art semi-supervised method. **ave**, **inter** and **cover** refers to average, intersection volume and contact coverage, respectively.

Methods	Hand error		Object ADD-0.1D(↑)				Contact	
	joint ↓	mesh ↓	bottle	can	bleach	avg	cover ↑	inter ↓
Initial Pose [16]	11.1	11.0	-	-	-	74.5	4.4	15.3±21.1
ContactOpt [12]	9.7	9.7	-	-	-	75.5	14.7	6.0±6.7
Liu <i>et al.</i> [31]	9.9	9.5	69.6	53.2	86.9	69.9	-	-
Ours	8.7	8.9	79.1	71.8	93.3	81.4	19.2	3.5±1.8

to generate pseudo-labels for hand. In contrast, we leverage physical contact and visual consistency constraints to generate pseudo contact labels which can be optimised jointly with hand and object poses. As shown in Table 2, our method outperforms [31] by 11.5% in average object ADD-0.1D score. Besides, we also compare with our baseline contact model ContactOpt [12]. As shown in Table 2, we are able to further improve hand error by 1mm and 0.8mm over joints and mesh. In addition to hand-object pose performance, our method is able to better reconstruct hand and object with less intersection volume and higher contact coverage. The above demonstrates that our method provides a more practical alternative to alleviate the reliance on heavy dataset annotation in hand-object. In addition, we provide qualitative comparison on *HO-3D* in Figure 5. We also report the cross-dataset generalisation performance of our model on *DexYCB* in Table 3. We select three objects (*i.e.*, mustard bottle, potted meat can and bleach cleanser), to be consistent with *HO-3D*. As shown, our method consistently shown improvements across all metrics.

Table 3. Error rates of the cross-dataset generalisation performance on *DexYCB*.

models	Hand error		Object ADD-0.1D($\uparrow$)				Contact	
	joint $\downarrow$	mesh $\downarrow$	bottle	can	bleach	avg	cover $\uparrow$	inter $\downarrow$
w/o semi-supervised	13.0	13.7	76.9	46.2	68.4	63.8	4.1	16.0 $\pm$ 12.3
semi-supervised	11.8	12.1	83.6	53.7	74.2	70.5	9.3	10.5$\pm$7.9

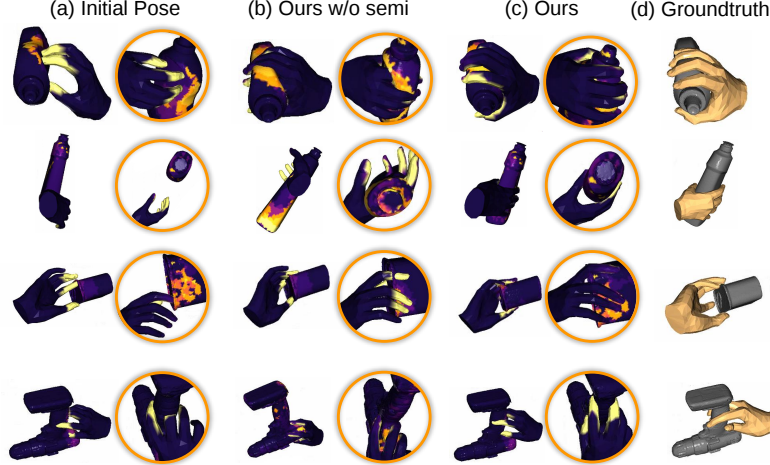


Fig. 5. Qualitative comparison with Initial Pose [31] on *HO-3D*. (a) We observed that the reconstructions of hand and object are more likely to be distant in 3D space due to depth and scale ambiguity with only RGB image as an input. (b) In contrast, when without semi-supervised learning, there exists more interpenetration or incorrect grasps between hand and object. (c) As shown, together with visual cues and contact map estimations, our method is able correct the above failure cases and generate more physically-plausible reconstructions.

4.2 Ablation study

Number of K neighbours. Table 4(a) shows the results of varying number of K neighbours without dilation. As shown, increasing K improves immediately at $K = 10$ but does not gain performance further.

Size of dilation factor d . As shown above that performance saturates at $K = 10$, we now fix K and vary the size of dilation factor d in Table 4(b). We find that the combination $K = 10, d = 4$ produce the best performance and do not improve by further increasing d . This demonstrates the effectiveness of increasing the receptive field for contact map estimation.

Combining K -NN computation. To further study the effect of separately computing K -NN, we experiment with combined K -NN computation with $d = 4$ in Table 4(c). It can be seen that the performance exceed the lower bound of (a) and similar to DGCNN’s performance in Table 1. This is expected as this is

Table 4. Performances of different GCN-Contact design choices on *ContactPose* and *HO-3D*. **semi** refers to semi-supervised learning. We experiment on (a) number of K neighbours without dilation, (b) size of dilation factor d with $K = 10$ and (c) combining K -NN computation (denoted with $*$) with $d = 4$. Full results are in supplementary.

models	<i>ContactPose</i> Hand error		<i>HO-3D</i> w/o semi		
	joint ↓	mesh ↓	joint ↓	mesh ↓	Object error add-0.1d (↑)
$K = 5$	8.252	8.134	12.69	12.86	66.33
(a) $K = 10$	6.691	6.562	11.81	11.91	68.71
$K = 15$	6.715	6.617	11.85	11.92	68.70
$d = 2$	5.959	5.865	10.91	10.86	70.25
(b) $d = 4$	5.878	5.765	9.92	9.79	72.81
$d = 6$	5.911	5.805	9.95	9.79	72.81
$K^* = 5$	8.451	8.369	12.91	12.86	68.86
(c) $K^* = 10$	8.359	8.251	11.55	11.97	69.10
$K^* = 25$	8.369	8.286	11.57	11.97	69.06

similar to static EdgeConv [55] with dilation. This shows that separate K -NN is crucial for this framework.

Impact of our components. We study the impact of semi-supervised learning on *HO-3D*. Since the hand model (MANO) is consistent across datasets, the contact estimator can easily transfer hand contact to new datasets without re-training. However, it is insufficient to adapt to unlabelled dataset due to diverse object geometries. Therefore, we propose a semi-supervised learning method to generate high-quality pseudo-labels. As shown in Table 5, our method enables performance boost on both hand and object. The hand joint error is $8.74mm$ while it is $9.92mm$ without semi-supervised training. Also, the average object ADD-0.1D has a significant 8.56% improvement under S²Contact.

Table 5 shows a quantitative comparison of S²Contact with various filtering constraints disabled demonstrating that constraints from both visual and geometry domains are essential for faithful training. We also observed that disabling $\mathcal{L}_{cont}$ can easily lead to unstable training and 5.45% performance degradation in object error. In contrast, geometric consistencies ($\mathcal{L}_{Cham}$ and $\mathcal{L}_{SDF}$) have a comparably smaller impact on hand and object pose. Despite that they account for less than 2% performance drop to object error, geometric consistencies are important for contact (*i.e.*, more than 5% for contact coverage). The remaining factor, measuring visual similarity, has a more significant impact. Disabling visual consistency constraint $\mathcal{L}_{SSIM}$ results in hand joint error and object error increase by $1mm$ and 8.04%, respectively. We validate that the combination of our pseudo-label filtering constraints are critical for generating high-quality pseudo-labels and improving hand-object pose estimation performance. Finally, we provide qualitative examples on out-of-domain objects in supplementary.

Table 5. Performances of different filtering constraints under semi-supervised learning on *HO-3D*. **semi** refers to semi-supervised learning.

models	Hand error		Object error add-0.1d ($\uparrow$)	Contact	
	joint $\downarrow$	mesh $\downarrow$		cover $\uparrow$	inter $\downarrow$
w/o semi	9.92	9.79	72.81	12.1	8.3 $\pm$ 10.5
w/ semi	8.74	8.86	81.37	19.2	3.5$\pm$1.8
w/o $\mathcal{L}_{Cham}$	8.53	8.48	80.11	13.1	6.9 $\pm$ 6.2
w/o $\mathcal{L}_{SDF}$	8.61	8.59	80.90	14.7	5.5 $\pm$ 6.0
w/o $\mathcal{L}_{cont}$	9.28	9.19	75.92	13.9	4.8 $\pm$ 3.1
w/o $\mathcal{L}_{SSIM}$	9.71	9.57	73.33	16.3	3.7 $\pm$ 2.6

Computational analysis. We report the model parameters and GPU memory cost in Table 4 of supplementary material. For fair comparisons, all models are tested using a batch size of 64. As shown, our model has 2.4X less the number of learnable model parameters and 2X less the GPU memory cost when compared to baseline and DGCNN [55], respectively. We alleviate the need to keep a high density of points across the network (DGCNN) while gaining performance.

5 Conclusion

In this paper, we have proposed a novel semi-supervised learning framework which enables learning contact with monocular videos. The main idea behind this study was to demonstrate that this can successfully be achieved with visual and geometric consistency constraints for pseudo-label generation. We designed an efficient graph-based network for inferring contact maps and shown benefits of combining visual cues and contact consistency constraints to produce more physically-plausible reconstructions. In the future, we would like to explore more consistencies over time and or multiple views to further improve the accuracy.

Acknowledgements. This research was supported by the MSIT (Ministry of Science and ICT), Korea, under the ITRC (Information Technology Research Center) support program (IITP-2022-2020-0-01789) supervised by the IITP (Institute of Information & Communications Technology Planning & Evaluation) and the Baskerville Tier 2 HPC service (<https://www.baskerville.ac.uk/>) funded by the Engineering and Physical Sciences Research Council (EPSRC) and UKRI through the World Class Labs scheme (EP/T022221/1) and the Digital Research Infrastructure programme (EP/W032244/1) operated by Advanced Research Computing at the University of Birmingham. KIK was supported by the National Research Foundation of Korea (NRF) grant (No. 2021R1A2C2012195) and IITP grants (IITP-2021-0-02068 and IITP-2020-0-01336). ZQZ was supported by China Scholarship Council (CSC) Grant No. 202208060266. AL was supported in part by the EPSRC (grant number EP/S032487/1). FZ was supported by the National Natural Science Foundation of China under Grant No. 61972188 and 62122035.

References

1. Brahmbhatt, S., Ham, C., Kemp, C.C., Hays, J.: ContactDB: Analyzing and predicting grasp contact via thermal imaging. In: CVPR (2019)
2. Brahmbhatt, S., Tang, C., Twigg, C.D., Kemp, C.C., Hays, J.: ContactPose: A dataset of grasps with object contact and hand pose. In: ECCV (2020)
3. Cao, Z., Radosavovic, I., Kanazawa, A., Malik, J.: Reconstructing hand-object interactions in the wild. In: ICCV (2021)
4. Chao, Y.W., Yang, W., Xiang, Y., Molchanov, P., Handa, A., Tremblay, J., Narang, Y.S., Van Wyk, K., Iqbal, U., Birchfield, S., et al.: DexYCB: A benchmark for capturing hand grasping of objects. In: CVPR (2021)
5. Chen, W., Jia, X., Chang, H.J., Duan, J., Leonardis, A.: G2L-Net: Global to local network for real-time 6D pose estimation with embedding vector features. In: CVPR (2020)
6. Chen, W., Jia, X., Chang, H.J., Duan, J., Shen, L., Leonardis, A.: FS-Net: Fast shape-based network for category-level 6D object pose estimation with decoupled rotation mechanism. In: CVPR (2021)
7. Chen, Y., Tu, Z., Ge, L., Zhang, D., Chen, R., Yuan, J.: SO-HandNet: Self-organizing network for 3D hand pose estimation with semi-supervised learning. In: CVPR (2019)
8. Corona, E., Pumarola, A., Alenya, G., Moreno-Noguer, F., Rogez, G.: GanHand: Predicting human grasp affordances in multi-object scenes. In: CVPR (2020)
9. Defferrard, M., Bresson, X., Vandergheynst, P.: Convolutional neural networks on graphs with fast localized spectral filtering. In: NeurIPS (2016)
10. Doosti, B., Naha, S., Mirbagheri, M., Crandall, D.J.: HOPE-Net: A graph-based model for hand-object pose estimation. In: CVPR (2020)
11. Garcia-Hernando, G., Yuan, S., Baek, S., Kim, T.K.: First-person hand action benchmark with RGB-D videos and 3D hand pose annotations. In: CVPR (2018)
12. Grady, P., Tang, C., Twigg, C.D., Vo, M., Brahmbhatt, S., Kemp, C.C.: ContactOpt: Optimizing contact to improve grasps. In: CVPR (2021)
13. Guo, M.H., Cai, J.X., Liu, Z.N., Mu, T.J., Martin, R.R., Hu, S.M.: PCT: Point cloud transformer. *Computational Visual Media* (2021)
14. Hampali, S., Rad, M., Oberweger, M., Lepetit, V.: Honnotate: A method for 3D annotation of hand and object poses. In: CVPR (2020)
15. Han, S., Liu, B., Cabezas, R., Twigg, C.D., Zhang, P., Petkau, J., Yu, T.H., Tai, C.J., Akbay, M., Wang, Z., et al.: MEgATrack: Monochrome egocentric articulated hand-tracking for virtual reality. In: SIGGRAPH (2020)
16. Hasson, Y., Tekin, B., Bogo, F., Laptev, I., Pollefeys, M., Schmid, C.: Leveraging photometric consistency over time for sparsely supervised hand-object reconstruction. In: CVPR (2020)
17. Hasson, Y., Varol, G., Laptev, I., Schmid, C.: Towards unconstrained joint hand-object reconstruction from RGB videos. In: 3DV (2021)
18. Hasson, Y., Varol, G., Tzionas, D., Kalevatykh, I., Black, M.J., Laptev, I., Schmid, C.: Learning joint reconstruction of hands and manipulated objects. In: CVPR (2019)
19. Huang, L., Tan, J., Meng, J., Liu, J., Yuan, J.: HOT-Net: Non-autoregressive transformer for 3D hand-object pose estimation. In: ACM MM (2020)
20. Jiang, H., Liu, S., Wang, J., Wang, X.: Hand-object contact consistency reasoning for human grasps generation. In: ICCV (2021)

21. Karunratanakul, K., Yang, J., Zhang, Y., Black, M.J., Muandet, K., Tang, S.: Grasping field: Learning implicit representations for human grasps. In: 3DV (2020)
22. Kato, H., Ushiku, Y., Harada, T.: Neural 3D mesh renderer. In: CVPR (2018)
23. Kaviani, S., Rahimi, A., Hartley, R.: Semi-Supervised 3D hand shape and pose estimation with label propagation. arXiv preprint arXiv:2111.15199 (2021)
24. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: ICLR (2015)
25. Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks. In: ICLR (2017)
26. Kwon, T., Tekin, B., Stühmer, J., Bogó, F., Pollefeys, M.: H2O: Two hands manipulating objects for first person interaction recognition. In: ICCV (2021)
27. Labbé, Y., Carpentier, J., Aubry, M., Sivic, J.: CosyPose: Consistent multi-view multi-object 6D pose estimation. In: ECCV (2020)
28. Li, G., Muller, M., Thabet, A., Ghanem, B.: DeepGNSs: Can GCNs go as deep as CNNs? In: ICCV (2019)
29. Li, Y., Wang, G., Ji, X., Xiang, Y., Fox, D.: DeepIM: Deep iterative matching for 6D pose estimation. In: ECCV (2018)
30. Lin, Z.H., Huang, S.Y., Wang, Y.C.F.: Convolution in the cloud: Learning deformable kernels in 3D graph convolution networks for point cloud analysis. In: CVPR (2020)
31. Liu, S., Jiang, H., Xu, J., Liu, S., Wang, X.: Semi-supervised 3D hand-object poses estimation with interactions in time. In: CVPR (2021)
32. Liu, Z., Hu, H., Cao, Y., Zhang, Z., Tong, X.: A closer look at local aggregation operators in point cloud analysis. In: ECCV (2020)
33. Maturana, D., Scherer, S.: VoxNet: A 3D convolutional neural network for real-time object recognition. In: IROS (2015)
34. Monti, F., Boscaini, D., Masci, J., Rodola, E., Svoboda, J., Bronstein, M.M.: Geometric deep learning on graphs and manifolds using mixture model CNNs. In: CVPR (2017)
35. Mueller, F., Bernard, F., Sotnychenko, O., Mehta, D., Sridhar, S., Casas, D., Theobalt, C.: GANerated hands for real-time 3D hand tracking from monocular RGB. In: CVPR (2018)
36. Mueller, F., Davis, M., Bernard, F., Sotnychenko, O., Verschoor, M., Otaduy, M.A., Casas, D., Theobalt, C.: Real-time pose and shape reconstruction of two interacting hands with a single depth camera. In: SIGGRAPH (2019)
37. Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., Devito, Z., Lin, Z., Desmaison, A., Antiga, L., Lerer, A.: Automatic Differentiation in Pytorch. In: NeurIPS (2017)
38. Qi, C.R., Su, H., Mo, K., Guibas, L.J.: PointNet: Deep learning on point sets for 3D classification and segmentation. In: CVPR (2017)
39. Qi, C.R., Yi, L., Su, H., Guibas, L.J.: PointNet++: Deep hierarchical feature learning on point sets in a metric space. In: NeurIPS (2017)
40. Qian, G., Hammoud, H., Li, G., Thabet, A., Ghanem, B.: ASSANet: An anisotropic separable set abstraction for efficient point cloud representation learning. NeurIPS (2021)
41. Romero, J., Tzionas, D., Black, M.J.: Embodied hands: Modeling and capturing hands and bodies together. ACM Transactions on Graphics (ToG) (2017)
42. Simon, T., Joo, H., Matthews, I., Sheikh, Y.: Hand keypoint detection in single images using multiview bootstrapping. In: CVPR (2017)

43. Spurr, A., Molchanov, P., Iqbal, U., Kautz, J., Hilliges, O.: Adversarial motion modelling helps semi-supervised hand pose estimation. arXiv preprint arXiv:2106.05954 (2021)
44. Spurr, A., Song, J., Park, S., Hilliges, O.: Cross-modal deep variational hand pose estimation. In: CVPR (2018)
45. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: CVPR (2015)
46. Taheri, O., Ghorbani, N., Black, M.J., Tzionas, D.: GRAB: A dataset of whole-body human grasping of objects. In: ECCV (2020)
47. Tang, D., Chang, H.J., Tejani, A., Kim, T.K.: Latent regression forest: Structured estimation of 3D articulated hand posture. In: CVPR (2014)
48. Tang, D., Yu, T.H., Kim, T.K.: Real-time articulated hand pose estimation using semi-supervised transductive regression forests. In: ICCV (2013)
49. Tekin, B., Bogó, F., Pollefeys, M.: H+O: Unified egocentric recognition of 3D hand-object poses and interactions. In: CVPR (2019)
50. Tse, T.H.E., De Martini, D., Marchegiani, L.: No need to scream: Robust sound-based speaker localisation in challenging scenarios. In: ICSR (2019)
51. Tse, T.H.E., Kim, K.I., Leonardis, A., Chang, H.J.: Collaborative learning for hand and object reconstruction with attention-guided graph convolution. In: CVPR (2022)
52. Ueda, E., Matsumoto, Y., Imai, M., Ogasawara, T.: A hand-pose estimation for vision-based human interfaces. *IEEE Transactions on Industrial Electronics* (2003)
53. Wang, H., Cong, Y., Litany, O., Gao, Y., Guibas, L.J.: 3DioUMatch: Leveraging IoU prediction for semi-supervised 3D object detection. In: CVPR (2021)
54. Wang, J., Mueller, F., Bernard, F., Sorli, S., Sotnychenko, O., Qian, N., Otaduy, M.A., Casas, D., Theobalt, C.: RGB2Hands: real-time tracking of 3D hand interactions from monocular RGB video. In: SIGGRAPH (2020)
55. Wang, Y., Sun, Y., Liu, Z., Sarma, S.E., Bronstein, M.M., Solomon, J.M.: Dynamic graph CNN for learning on point clouds. In: SIGGRAPH (2019)
56. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing* **13**(4), 600–612 (2004)
57. Wu, W., Qi, Z., Fuxin, L.: PointConv: Deep convolutional networks on 3D point clouds. In: CVPR (2019)
58. Xiang, Y., Schmidt, T., Narayanan, V., Fox, D.: PoseCNN: A convolutional neural network for 6D object pose estimation in cluttered scenes. In: RSS (2018)
59. Xu, M., Ding, R., Zhao, H., Qi, X.: PAConv: Position adaptive convolution with dynamic kernel assembling on point clouds. In: CVPR (2021)
60. Yang, J., Chang, H.J., Lee, S., Kwak, N.: SeqHAND: RGB-sequence-based 3D hand pose and shape estimation. In: ECCV (2020)
61. Yang, L., Chen, S., Yao, A.: SemiHand: Semi-supervised hand pose estimation with consistency. In: ICCV (2021)
62. Yang, L., Zhan, X., Li, K., Xu, W., Li, J., Lu, C.: CPF: Learning a contact potential field to model the hand-object interaction. In: ICCV (2021)
63. You, H., Feng, Y., Ji, R., Gao, Y.: PVNet: A joint convolutional network of point cloud and multi-view for 3D shape recognition. In: ACM Multimedia (2018)
64. Zhang, T., McCarthy, Z., Jow, O., Lee, D., Chen, X., Goldberg, K., Abbeel, P.: Deep imitation learning for complex manipulation tasks from virtual reality teleoperation. In: ICRA (2018)
65. Zhao, H., Jiang, L., Jia, J., Torr, P.H., Koltun, V.: Point transformer. In: ICCV (2021)

- 66. Zheng, L., Leonardis, A., Tse, T.H.E., Horanyi, N., Chen, H., Zhang, W., Chang, H.J.: TP-AE: Temporally primed 6D object pose tracking with auto-encoders. In: ICRA (2022)
- 67. Zimmermann, C., Brox, T.: Learning to estimate 3D hand pose from single RGB images. In: ICCV (2017)