Hand-Object Pose Estimation from RGBD Images

Anonymous CVPR 2023 submission

Paper ID

Abstract

Hand-object pose estimation aims to predict the pose and shape of both the hand and held object under interaction. Although numerous applications in the real world such as augmented and virtual reality, hand-object pose estimation concerns relatively less attention. Several methods separately estimate hand shapes and object poses but totally neglect the correlations between hands and objects. In this work, we introduce an approach that leverages the advantage of the voting mechanism to jointly learn the appearance of hands and objects from RGB-D images. We propose a module called adaptive fusion to effectively collaborate RGB and Depth features by adaptively adding the weight for each type of feature before fusing. The output features can discriminate the differently meaningful distribution between color and depth information at each position. Moreover, we embrace the graph convolutional network (GCN) to learn the interaction relationships between the hand and held object shapes under manipulation. Unlike conventional methods looking at the center points of the hand and objects, our approach takes into account the hand and object keypoints that belong to their surfaces to estimate the shapes and learn the correlations. This facilitates examining the geometric constraints and spatial restrictions to achieve accurate outcomes. Experiments using benchmark datasets illustrate that our network achieves beyond state-of-the-art accuracy in 3D pose estimation.

1. Introduction

Estimation of hands and objects is fundamental and crucial for understanding meaningful interpretation of human action and behavior. It provides enormous knowledge for environmental perception and teaching manipulating systems. With the advent of deep learning, pose estimation tasks have significantly made progress such as RGB-based [3,9,36,42,49], depth-based [2,10,19,23,24,43], and RGB-D methods [16,45]. Jointly estimation of hands and objects under interaction, however, has attracted less attention due to chronic challenges. This requires simultaneously predicting the pose and shape of hands and objects during the hand manipulating the objects. In this paper, we propose a novel network to tackle this problem from RGB-D images.

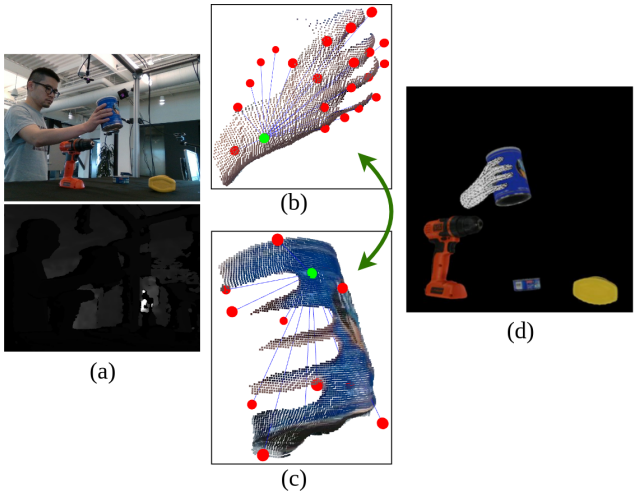

Figure 1: Our method captures features from both RGB and depth images (a) to vote for hand and object keypoints. (b) The hand point cloud and voting vectors (blue lines), each point on the hand (green point) votes for 21 MANO joints (red points). (c) The held object and voting vectors (blue lines), each point on the same object (green point) votes for $M = 9$ selected keypoints (red points). The interaction relationships between the hand and the held object are learn by feeding voting vectors into a GCN. (d) The hand-object pose estimation result.

Joint hand-object pose estimation under interaction is a much more challenging problem. The hand shape is notorious for being self-occlusion. This problem is adversely serious in the context when the hand manipulating an object due to escalated occlusion. The naive approach is estimating the shape of the hands and objects seperately. Such methods leverage the success of object pose estimation and hand shape reconstruction independently without considering the correlations between themselves. They totally ig-

nore the heavy interdependencies between hand pose on the held object's shape. Intuitively, the presence of object strongly defines and constrains the hand grasps and therefore limits the feasible hand gestures to a restricted number. Similarly, determining hand gestures provides a cue for estimating the shape and pose of the held object. Consequently, simultaneously predicting the shapes of both hand and object soonly catches the attention of computer vision researchers. [12–14, 20, 21, 26, 34, 37, 41] joinly learn the hand and object poses from a single RGB image. Whereas, [6, 11, 25, 47] focus on another input format, depth images, to achieve the expected results. However, they pose a threat to the prediction accuracy due to lack of the other type format input. With the prevalence of depth sensor, RGB-D image-based methods [18, 38] appear to be a promising solution. Although numerous research has made an impressive success in a wide range of computer vision tasks. It still puzzles the reseacher community of how to effectively using RGB-D input for joint hand-object pose estimation.

In this paper, we propose a network termed adaptive fusion that extracts color and depth features and wisely combines them to generate the discriminative representation of input data. We embrace the fact that the favourable features conveyed by color and depth information differ across positions. At a specific pixel, the RGB feature may be much more compelling than the physical one but the other pixel may witness the opposite situation. Our network learns to predict the weight factors that either amplifies the favourable features or inhibits the meaningless information. Therefore, it can thoroughly investiagte the meaningfulness and the beneficial contribution of RGB and Depth values at each position to effectively incorporate them together across all positions inspired by [40]. In other words, our proposal module does not solely integrate the geometric feature to the color one, but also tells to what extent the system should pay attention on each type of features at each position.

To tackle the problem of adverse occlusion under the hand-object interaction contexts, we adopt the graph convolutional network (GCN) to introduce the interaction learning procedure that captures the correlations between hands and objects appearance. We firstly predict the hand and object keypoints by deep a Hough voting network. The network focuses on the keypoints of hands and objects instead of their centroids, which are dominant in detection and pose estimation tasks. This is because the centroid points are capable to represent the object's existence but cannot provide the information for hand-object interaction learning due to their properties of being away from the contacting surfaces and being virtual. The keypoints, hence, are favourable to GCN to absorb the geometric interdependencies and constraints between hand and object shapes. This facilitates our network can consider the shape-defining conditions that the hand puts on the manipulated object and in the reverse relation. Therefore, this boosts the pose estimation performance of our method when coping with the interaction contexts. We further conduct experiments on **dataset** to evaluate our method. Experimental results demonstrate that our approach performances significantly exceed the current state-of-the-art methods.

In brief, the main contributions of our work are:

- We propose the adaptive fusion network to empower the capability of extracting features from RGB-D images. This network can learn both color and geometric features and then adaptively fuse them by either wisely and selectively magnifying the value features or weakening the useless one at each pixel.
- We introduce a GCN-based network to learn the interrelation between the hand and the manipulated object poses by examining their corresponding voting vectors. It allows to learn the hand pose under the impact of the object's existence and vice versa.
- Experiments on benchmark datasets demonstrate that our approach can outweigh the state-of-the-art models for hand and object 3D pose estimation.

2. Related work

2.1. Hand-object pose estimation

The naive approach for this mission is treating the hand [7, 15, 22, 24, 44] and manipulated object [28, 33, 40, 48] separately without considering their interdependence. They underestimate the extraordinary relationship between hand gestures and object shapes. Several approaches overcome this problem by jointly learning the shapes of both hands and objects from RGB images. [12] looks into photometric consistency between neighboring frames to reconstruct hand-object shapes under interactions. [34] handles hand action classification to assist the process of estimating hand-object interactions. [20] introduce a semi-supervised learning framework leveraging spatial-temporal consistency to improve estimation performance. However, the absence of depth information makes the process of learning physical constraints and interactions latent. In addition, the transformation from 2D to a 3D world is accurately difficult due to the high degree of nonlinearity. In contrast, some methods solely exploit depth images. [6] designs an architecture that firstly predicts hand and object centers and then learns global orientations and hand grasp configurations while interacting with objects. [11, 47] segments 2D hand and object regions from depth images and then optimizes the reconstruction process of interacting motions. This method, however, learns the depth images by a 2D CNN backbone and hence cannot radically observe the geometric information. [25] deploys a feedback loop to revise the flawed estimation results using depth images only. RGB-D input, on

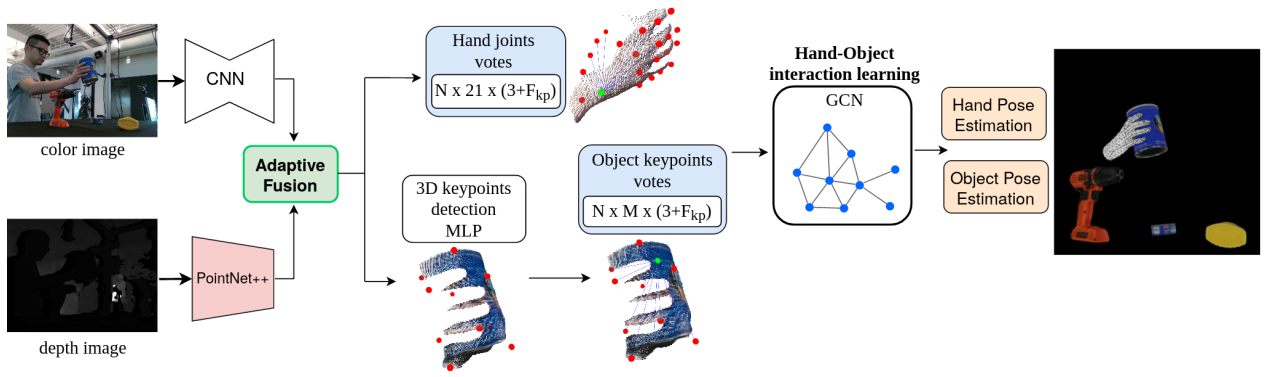

Figure 2: **Overview of our proposal**. Our method takes both color and depth maps as input data. The color features are extracted by a CNN, while the 3D features are calculated by PointNet++ architecture. These two types of features are then fused at a pixel level to obtain the new distinctive features by the adaptive fusion network. This network learns to predict the weight matrix that facilitates beneficial features to eclipse the tedious information. We design a deep Hough voting-based network to vote for 21 MANO hand joints and $M = 9$ selected object keypoints. A GCN-based network is deployed to learn the constraints and interdependencies between the hand and object poses to boost the estimation performance.

the other hand, has received relatively less attention due to the puzzle of an effective collaboration of two distinctive input formats still holds a secret. [18] focuses on the physical laws of hand actions from RGB-D input to benefit hand-object interaction interpretations. [38] tracks hands and objects in dealing with a complex scenario in which manipulated objects are deformable.

2.2. RGB-D fusion

With the common of color-depth camera, a wide range of computer vision research such as object segmentation [4, 5, 27, 46] and 6D object detection [32, 35, 40] has been inspired to learn and incorporate color and depth features from RGB-D images. The RGB image and depth image belong to different modalities, so most fusing feature methods are [39]: image layer fusion, feature layer fusion, and output layer fusion. While image layer fusion concatenates the input data before feeding to CNNs, feature layer fusion means learning color and depth data in two distinguished architectures but sharing the learning process. Output layer fusion, on the other hand, integrates two feature maps that are separately extracted by two backbone networks. However, fusion RGB-D features for hand-object pose estimation is less attractive because most of the mentioned methods have a mutual weak point which is extracting features from depth maps by 2D CNNs. This makes the 3D spatial feature output latent and oblivious. Motivated by [40], we develop a network that not only exports geometric information and geometric constraints by using Pointnet++ [30], but also adaptively and selectively adjusts features at each pixel before fusing to achieve the reliable performance.

2.3. Graph convolutional network for pose estimation

The power of graph convolutional networks has recently captured the researchers' attention in solving the problem of pose estimation. The network provides the relationship awareness of input data, hence can cope with the problem of a large number of DoF in hand pose estimation. [17] allows the network to be spatially aware to boost the 2D hand keypoints prediction. Interm of jointly estimate hand-object pose estimation, [8] develops two graph convolutional network-based architectures for two missions. The first one detects 2D hand joints and 2D object corners, while the second lifts 2D keypoints to 3D coordinates. [1] also designs two steps of firstly predicting 2D hand-object poses. Afterward, this method boosts the performance by gradually providing more information such as the third dimension and mesh vertices to the graph-based network. [37] proposes attention-guided graph convolution to iteratively share hand and object estimator between two branches for learning the mutual occlusion. This success inspires us to develop a graph-based network for hand-object interaction learning but instead directly feed the 3D features to learn the relationship without the need of lifting 2D information to 3D coordinates, which might cause unpredictable mistakes.

3. Method

In Figure 2, we provide an overall pipeline of our framework for hand-object pose estimation under interaction. The following details our work.

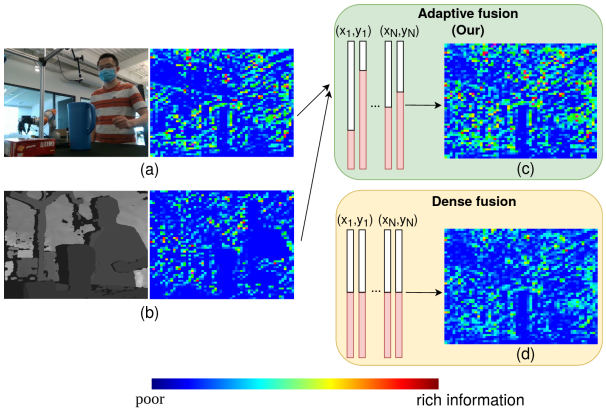

Figure 3: **Adaptive fusion.** Illustration of our proposal adaptive network in comparison with Dense fusion [40]. The network takes color feature (a) and depth feature (b) as input and learns their diverse distribution of contributing meaningfulness across all positions. Whereas, Dense fusion values RGB and depth feature at each pixel equally. The white and pink columns' height denotes the important degree of each type of feature at each pixel. Intuitively, our output feature (c) is more densely informative than dense fusion output (d).

3.1. Adaptive Fusion

Dense fusion [40] has recently made a remarkable stride in effectively exploiting color and depth images due to two main contributions. The first merit is the capability of extracting geometric features, which conventional RGB-D fusion methods fail to obtain due to applying 2D CNNs for depth images. The second one is introducing pixel-wise fusion to concatenate color and geometric features at a pixel level. However, the network believes 2D and 3D information carried by each pixel are equally valued, while the fact is the opposite. At a specific position, the 2D feature may be remarkable, while the 3D information is latent. To tackle this problem, we propose an adaptive fusion network that learns the favourability distribution of 2D and 3D features for the hand-object pose estimation performance. Figure 3 illustrates the difference between our work and Dense fusion and presents our more-appealing feature map outcome.

Color feature extraction: Given a color image $I_{rgb} \in \mathbb{R}^{H\times W\times 3}$, the color features $f_{rgb} = \{f_i^{rgb}\}_{i=1}^{H\times W}$ are normally extracted by a CNN architecture. Where $f_{rgb} \in \mathbb{R}^{H\times W\times d_{rgb}}$ and each pixel is mapped into a color feature space $f_i^{rgb} \in \mathbb{R}^{d_{rgb}}$.

Depth feature extraction: The geometric features, on the other hand, are extracted by converting depth maps to point cloud and then feeding into PointNet [29]. In our work, differing from the original work, we conduct PointNet++ [30], an upgraded version, to replace the original backbone. Given a depth map $I_d \in \mathbb{R}^{H\times W\times 1}$, the point cloud features $f_{geo} = \{f_i^{geo}\}_{i=1}^{H\times W}$.

Feature Embedding: To discriminate the favourability of color and depth features, we add we add learnable weighting matrixes before the fusion process. These matrixes allow each type of feature at each position to be either accelerated or vanished. The output feature is computed as equation 1, where A and B are learnable hyper-parameters. $f^{fusion} \in \mathbb{R}^{H\times W\times(d_{rgb}+d_{geo})}$.

$$f^{fusion} = A \times f^{rgb} \oplus B \times f^{geo} \tag{1}$$

3.2. Hand and Object Voting

Hand joints voting: As shown in 2, the discriminative features with rich information after the fusion procedure are used to regress hand joints. Conventional voting methods approach object pose estimation including hand poses usually vote for the hand center. Whereas, our method computes votes for hand joints points since hand joints can reflect the hand gestures, which is crucial for hand pose estimation under interaction. The hand joints convey information about the hand shape itself but also the 3D object shape. Therefore, such hand joints are necessary for hand-object interaction learning. We adopt the MANO hand mesh model [31] with 21 hand keypoints J consisting of 16 original hand joints and 5 hand vertices.

Given the point cloud $\{p_i\}_{i=1}^{N_\mathcal{H}}$ and 21 MANO hand keypoints $\{Hkp_j\}_{j=1}^{21}$ belong to the same hand $\mathcal{H}$. We denote $p_i = [x_i, f_i^{fusion}]$ with x_i the 3D coordinate and f_i^{fusion} the attentionally fused feature. Similarly, we denote $Hkp_j = [x_j^{Hkp}]$ with x_j^{Hkp} the 3D coordinate of the hand keypoints. We compute the translation offset $\{\Delta_{Hkp_i}^{\ j}\}_{j=1}^{21}$ for each point, where $\Delta_{Hkp_i}^{\ j}$ denotes the translation offset from the i_{th} point to the j_{th} hand keypoint. Thevoted keypoint can be computed as $vHkp_i^j = x_i + \Delta_{Hkp_i}^{\ j}$. We define the loss for hand keypoints learning as below:

$$\mathcal{L}_{Hkp} = \frac{1}{N_\mathcal{H}} \sum_{i=1}^{N_\mathcal{H}} \sum_{j=1}^{21} \|\Delta_{Hkp_i}^{\ j} - \Delta_{Hkp_i}^{\ j*}\|_H \cdot \mathbb{1}(p_i \in \mathcal{H}) \tag{2}$$

where $\Delta_{Hkp_i}^{\ j*}$ is the ground truth translation offset, $N_\mathcal{H}$ is the total number of points belonging to a hand $\mathcal{H}$. $\|\cdot\|_H$ is the Huber norm. The binary function $\mathbb{1}(\cdot)$ equals to 1 when point p_i belongs to a hand $\mathcal{H}$, and 0 otherwise.

Object keypoints Selection: The 3D keypoints are selected from 3D object models. Normally, eight corners of the 3D bounding box are used to represent the object [8, 17]. However, the corner points are actually far away from points on objects, leading to the difficulty to infer the physical constraints while interacting with the hand. Therefore, we instead select keypoints on the object surfaces that provide ease to learning the hand-object interaction. We

use the farthest point sampling (FPS) algorithm to collect the keypoints of objects by initializing an object mesh center point as the first keypoint and then searching the others by FPS until obtaining M keypoints.

Object keypoints voting: In terms of learning the object presence, the attentionally fused features are fed into a module to predict 3D keypoints for each object. Concretely, given a set of points $\{p_i\}_{i=1}^{N_\mathcal{O}}$ and M selected object keypoints $\{Okp_j\}_{j=1}^{M}$ belong to the same object $\mathcal{O}$. We denote $Okp_j = [x_j^{Okp}]$ with x_j^{Okp} the 3D coordinate of the object keypoints. The translation offset from the i_th point to the j_th object keypoints is denoted as $\Delta_{Okp_i^j}$. Hence, for each point we generate translation offset $\{\Delta_{Hkp_i^j}\}_{j=1}^{M}$. The voted object keypoint can be computed as $vOkp_i^j = x_i + \Delta_{Okp_i^j}$. We define the loss function as below:

$$\mathcal{L}_{Okp} = \frac{1}{N_\mathcal{O}} \sum_{i=1}^{N_\mathcal{O}} \sum_{j=1}^{M} \|\Delta_{Okp_i^j} - \Delta_{Okp_i^{j*}}\|_H \cdot \mathbb{1}(p_i \in \mathcal{O}) \tag{3}$$

where $\Delta_{Okp_i^{j*}}$ is the ground truth translation offset, $N_\mathcal{O}$ is the total number of points belonging to an object $\mathcal{O}$. $\|\cdot\|_H$ is the Huber norm. The binary function $\mathbb{1}(\cdot)$ equals to 1 when point p_i belongs to an object $\mathcal{O}$, and 0 otherwise.

3.3. Hand and Object Poses Estimation

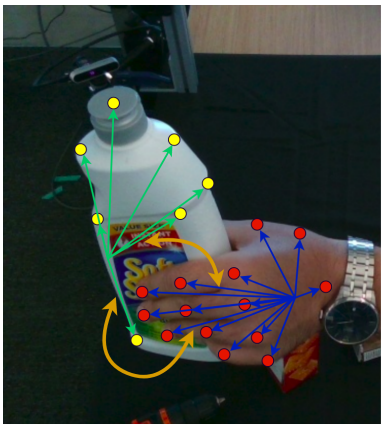

Figure 4: Illustration of the interaction learning between votes for hand keypoints and votes for object keypoints. The red points denote hand keypoints, while the yellow ones denote object keypoints. The blue and green vectors represent hand keypoints and object keypoints votes, respectively.

Hand-object interaction learning: To estimate the hand and object shapes under interactions, voting vectors should be aware of their global neighborhood. Especially, the object keypoints in the vicinity are intuitively beneficial for predicting hand keypoints and vice versa. We adopt a graph convolutional network (GCN) for the interaction learning procedure. Each node of the graph is defined by the proposal position y_i associated with proposal feature g_i. In particular, the proposal position is either hand keypoint ($y_i = vHkp_i^j$) or object keypoint ($y_i = vOkp_i^j$) and the associated proposal feature $g_i = f_i^{fusion}$. An edge between two nodes is determined by checking the condition of the Euclidean distance between them. If the distance between two neighboring positions ($d_{y_i,y_j} < \delta$), the edge-feature is defined as:

$$e_{ij} = h([y_i, g_i], [y_j, g_j] - [y_j, g_j]) \tag{4}$$

where $\langle$ is a non-linear fuction. Obtain refined proposal features from intial fusion featues.

Hand and Object pose regression: we adopt the MANO hand mesh model defined as a manifold triangle mesh $M = (V, F)$ to estimate the final hand pose. $V = \{v_i \in \mathbb{R}^3\}|1 \leq i \leq n$ is a set of $n = 778$ vertices and F is a set of faces. They are parameterized by the MANO parameters ($\theta \in \mathbb{R}^{51}, \beta \in \mathbb{R}^{10}$). We use multi-layer perceptron (MLP) to regress the parameters (θ, β). We define the loss function for hand pose regression as below, where the hand keypoints loss $\mathcal{L}_{Hkp}$ as equation 2.

$$\mathcal{L}_{handpose} = \mathcal{L}_{Hkp} + \mathcal{L}_V + \mathcal{L}_\theta + \mathcal{L}_\beta \tag{5}$$

In terms of regressing the object pose, we embrace the procedure that maps 6D vectors in representation space produced by the network into the original rotation space and minimizes the differences between the output and the ground-truth rotation matrices. The rigid transformation consists of a rotation $R \in SO(3)$ and a translation $t \in \mathbb{R}^3$. We define the loss function as below:

$$\mathcal{L}_{objectpose} = \mathcal{L}_{Okp} + \mathcal{L}_t + \mathcal{L}_R \tag{6}$$

where the loss for object keypoints voting $\mathcal{L}_{Okp}$ is defined as equation 3, $\mathcal{L}_t$ is the translation loss. The above rotation loss $\mathcal{L}_R$ is appropriate to asymmetric objects. The rotation metric for symmetric objects is diverse, therefore, given the estimated rotation $\overline{R}$ and translation $\overline{t}$ and the ground-truth (R^*, t^*). The rotation loss is redefined as below:

$$\mathcal{L}_R = \frac{1}{m} \sum_{x_1 \in \mathcal{M}} \| \min_{x_2 \in \mathcal{M}} (\overline{R}x + \overline{t} - R^*x - t^*) \| \tag{7}$$

where $\mathcal{M}$ denotes the 3D object models and m is the number of points.

Finally, the loss function for hand-object pose estimation under interaction is summarized as below:

$$\mathcal{L}_{hand-object} = \mathcal{L}_{handpose} + \mathcal{L}_{objectpose} \tag{8}$$

4. Evaluation

References

[1] Murad Almadani, Ahmed Elhayek, Jameel Malik, and Didier Stricker. Graph-based hand-object meshes and poses reconstruction with multi-modal input. *IEEE Access*, 9:136438–136447, 2021. 3

[2] Dingding Cai, Janne Heikkilä, and Esa Rahtu. Ove6d: Object viewpoint encoding for depth-based 6d object pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6803–6813, 2022. 1

[3] Yujun Cai, Liuhao Ge, Jianfei Cai, Nadia Magnenat Thalmann, and Junsong Yuan. 3d hand pose estimation using synthetic data and weakly labeled rgb images. *IEEE transactions on pattern analysis and machine intelligence*, 43(11):3739–3753, 2020. 1

[4] Sihan Chen, Xinxin Zhu, Wei Liu, Xingjian He, and Jing Liu. Global-local propagation network for rgb-d semantic segmentation. *arXiv preprint arXiv:2101.10801*, 2021. 3

[5] Xiaokang Chen, Kwan-Yee Lin, Jingbo Wang, Wayne Wu, Chen Qian, Hongsheng Li, and Gang Zeng. Bi-directional cross-modality feature propagation with separation-and-aggregation gate for rgb-d semantic segmentation. In *European Conference on Computer Vision*, pages 561–577. Springer, 2020. 3

[6] Chiho Choi, Sang Ho Yoon, Chin-Ning Chen, and Karthik Ramani. Robust hand pose estimation during the interaction with an unknown object. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3123–3132, 2017. 2

[7] Xiaoming Deng, Shuo Yang, Yinda Zhang, Ping Tan, Liang Chang, and Hongan Wang. Hand3d: Hand pose estimation using 3d neural network. *arXiv preprint arXiv:1704.02224*, 2017. 2

[8] Bardia Doosti, Shujon Naha, Majid Mirbagheri, and David J Crandall. Hope-net: A graph-based model for hand-object pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6608–6617, 2020. 3, 4

[9] Yafei Gao, Yida Wang, Pietro Falco, Nassir Navab, and Federico Tombari. Variational object-aware 3-d hand pose from a single rgb image. *IEEE Robotics and Automation Letters*, 4(4):4239–4246, 2019. 1

[10] Liuhao Ge, Hui Liang, Junsong Yuan, and Daniel Thalmann. 3d convolutional neural networks for efficient and robust hand pose estimation from single depth images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1991–2000, 2017. 1

[11] Duncan Goudie and Aphrodite Galata. 3d hand-object pose estimation from depth with convolutional neural networks. In *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, pages 406–413. IEEE, 2017. 2

[12] Yana Hasson, Bugra Tekin, Federica Bogo, Ivan Laptev, Marc Pollefeys, and Cordelia Schmid. Leveraging photometric consistency over time for sparsely supervised hand-object reconstruction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 571–580, 2020. 2

[13] Yana Hasson, Gül Varol, Cordelia Schmid, and Ivan Laptev. Towards unconstrained joint hand-object reconstruction from rgb videos. In *2021 International Conference on 3D Vision (3DV)*, pages 659–668. IEEE, 2021. 2

[14] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11807–11816, 2019. 2

[15] Umar Iqbal, Pavlo Molchanov, Thomas Breuel Juergen Gall, and Jan Kautz. Hand pose estimation via latent 2.5 d heatmap regression. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 118–134, 2018. 2

[16] Evangelos Kazakos, Christophoros Nikou, and Ioannis A Kakadiaris. On the fusion of rgb and depth information for hand pose estimation. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 868–872. IEEE, 2018. 1

[17] Deying Kong, Haoyu Ma, and Xiaohui Xie. Sia-gcn: A spatial information aware graph neural network with 2d convolutions for hand pose estimation. *arXiv preprint arXiv:2009.12473*, 2020. 3, 4

[18] Nikolaos Kyriazis and Antonis Argyros. Physically plausible 3d scene tracking: The single actor hypothesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9–16, 2013. 2, 3

[19] 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. 1

[20] Shaowei Liu, Hanwen Jiang, Jiarui Xu, Sifei Liu, and Xiaolong Wang. Semi-supervised 3d hand-object poses estimation with interactions in time. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14687–14697, 2021. 2

[21] Yao Lu and Walterio W Mayol-Cuevas. Understanding egocentric hand-object interactions from hand pose estimation. *arXiv preprint arXiv:2109.14657*, 2021. 2

[22] Meysam Madadi, Sergio Escalera, Xavier Baró, and Jordi Gonzalez. End-to-end global to local cnn learning for hand pose recovery in depth data. *arXiv preprint arXiv:1705.09606*, 2017. 2

[23] 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. 1

[24] Markus Oberweger and Vincent Lepetit. Deepprior++: Improving fast and accurate 3d hand pose estimation. In *Proceedings of the IEEE international conference on computer vision Workshops*, pages 585–594, 2017. 1, 2

[25] Markus Oberweger, Paul Wohlhart, and Vincent Lepetit. Generalized feedback loop for joint hand-object pose estimation. *IEEE transactions on pattern analysis and machine intelligence*, 42(8):1898–1912, 2019. 2

[26] 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 *2011 International Conference on Computer Vision*, pages 2088–2095. IEEE, 2011. 2

[27] Seong-Jin Park, Ki-Sang Hong, and Seungyong Lee. Rdfnet: Rgb-d multi-level residual feature fusion for indoor semantic segmentation. In *Proceedings of the IEEE international conference on computer vision*, pages 4980–4989, 2017. 3

[28] Charles R Qi, Wei Liu, Chenxia Wu, Hao Su, and Leonidas J Guibas. Frustum pointnets for 3d object detection from rgb-d data. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 918–927, 2018. 2

[29] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 652–660, 2017. 4

[30] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, 30, 2017. 3, 4

[31] Javier Romero, Dimitrios Tzionas, and Michael J Black. Embodied hands: Modeling and capturing hands and bodies together. *arXiv preprint arXiv:2201.02610*, 2022. 4

[32] Lounes Saadi, Bassem Besbes, Sebastien Kramm, and Abdelaziz Bensrhair. Optimizing rgb-d fusion for accurate 6dof pose estimation. *IEEE Robotics and Automation Letters*, 6(2):2413–2420, 2021. 3

[33] 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. 2

[34] Bugra Tekin, Federica Bogo, and Marc Pollefeys. H+ o: Unified egocentric recognition of 3d hand-object poses and interactions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4511–4520, 2019. 2

[35] Meng Tian, Liang Pan, Marcelo H Ang, and Gim Hee Lee. Robust 6d object pose estimation by learning rgb-d features. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6218–6224. IEEE, 2020. 3

[36] Jonathan Tremblay, Thang To, Balakumar Sundaralingam, Yu Xiang, Dieter Fox, and Stan Birchfield. Deep object pose estimation for semantic robotic grasping of household objects. *arXiv preprint arXiv:1809.10790*, 2018. 1

[37] Tze Ho Elden Tse, Kwang In Kim, Ales Leonardis, and Hyung Jin Chang. Collaborative learning for hand and object reconstruction with attention-guided graph convolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1664–1674, 2022. 2, 3

[38] Aggeliki Tsoli and Antonis A Argyros. Joint 3d tracking of a deformable object in interaction with a hand. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 484–500, 2018. 2, 3

[39] Changshuo Wang, Chen Wang, Weijun Li, and Haining Wang. A brief survey on rgb-d semantic segmentation using deep learning. *Displays*, 70:102080, 2021. 3

[40] Chen Wang, Danfei Xu, Yuke Zhu, Roberto Martín-Martín, Cewu Lu, Li Fei-Fei, and Silvio Savarese. Densefusion: 6d object pose estimation by iterative dense fusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3343–3352, 2019. 2, 3, 4

[41] Tiancai Wang, Tong Yang, Martin Danelljan, Fahad Shahbaz Khan, Xiangyu Zhang, and Jian Sun. Learning human-object interaction detection using interaction points. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4116–4125, 2020. 2

[42] 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. 1

[43] Fu Xiong, Boshen Zhang, Yang Xiao, Zhiguo Cao, Taidong Yu, Joey Tianyi Zhou, and Junsong Yuan. A2j: Anchor-to-joint regression network for 3d articulated pose estimation from a single depth image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 793–802, 2019. 1

[44] John Yang, Yash Bhalgat, Simyung Chang, Fatih Porikli, and Nojun Kwak. Dynamic iterative refinement for efficient 3d hand pose estimation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1869–1879, 2022. 2

[45] Shanxin Yuan, Bjorn Stenger, and Tae-Kyun Kim. 3d hand pose estimation from rgb using privileged learning with depth data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019. 1

[46] Guodong Zhang, Jing-Hao Xue, Pengwei Xie, Sifan Yang, and Guijin Wang. Non-local aggregation for rgb-d semantic segmentation. *IEEE Signal Processing Letters*, 28:658–662, 2021. 3

[47] Hao Zhang, Yuxiao Zhou, Yifei Tian, Jun-Hai Yong, and Feng Xu. Single depth view based real-time reconstruction of hand-object interactions. *ACM Transactions on Graphics (TOG)*, 40(3):1–12, 2021. 2

[48] Yin Zhou and Oncel Tuzel. Voxelnet: End-to-end learning for point cloud based 3d object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4490–4499, 2018. 2

[49] Christian Zimmermann and Thomas Brox. Learning to estimate 3d hand pose from single rgb images. In *Proceedings of the IEEE international conference on computer vision*, pages 4903–4911, 2017. 1