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 of the hand and held object under interaction. Although having 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 neglecting the correlations between hands and objects. In this work, we propose an approach that leveraging the advantage of voting mechanism to jointly learn the appearance of hand and object from RGB-D images. Our method effectively collaborates RGB and Depth features by sharing and fusing them at pixel level during the extraction process. The output features discriminate the differently meaningful distribution between color and depth information at each position to generate discriminative representations of RGB-D input. Moreover, we introduce a network to collaboratively learn voting vectors for both of the hand and object appearances to estimate their poses. This facilitates our network to examine their constrains and interactions to produce 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 behaviour. 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 [2, 10, 40, 47, 54], depth-based [1, 11, 22, 26, 27, 48], and RGB-D methods [19, 50]. Jointly estimation of hands 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 hands handling and executing the objects. In this paper, we propose a novel network to tackle this problem from RGB-D images.

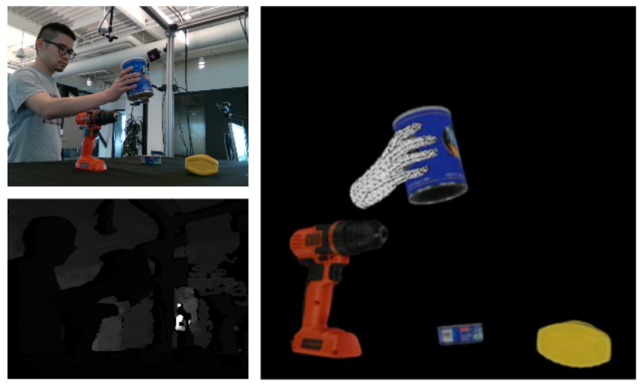

Figure 1: Example of RGB-D.

Joint hand-object pose estimation under interaction, on the other hand, 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. 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 ignore the heavily dependencies of hand pose on the held object's shape, and and vice versa. 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.

Inspired by the above perspective, some deep learning-based methods [13–15, 23, 24, 29, 37, 41, 45] joinly learn the hand and object poses from a single RGB image. Whereas, [5, 12, 28, 52] 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 camera, RGB-D image-based methods [21, 42] 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 that firstly extracts color and depth features and combines them to generate the discriminative representation of input data. The color feature is extracted by convolutional neural networks (CNNs), while the geometric information is learnt by PointNet++ [32]. The PointNet++ architecture empowers our framework to learn the physical constraints and geometric relationships between the hand and object, which are essential for estimating hand and object poses simultaneously. Differing [7], in which we find the inspiration, our architecture allows sharing information between two backbone networks. This helps the process of learning one type of input feature can absorb the presence of the other one. Therefore, our method can thoroughly investiagte the meaningfulness and the beneficial contribution of RGB and Depth values at each position across all positions. Furthermore, we develop a technique based on pixel-wise fusion [44] to attentionaly integrate geometric information into color features. 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. To handle this problem, our method introduces a learnable weight parameter to either facilitate or inhibit the feature at each pixel before fusing. In other words, our proposal network does not solely integrate the geometric feature to the color one at pixel level, but also tells to what extent the system should pay attention on each type of features at each position.

In terms of pose estimation, we adpot the voting mechanism to predict the hand and object poses simultaneously. The voting mechanism [7, 16, 46] has recently emerged as a compelling strategy for robustly forecasting the shapes. This is attributed to that voting methods can meet successful outcomes without requiring pre-known CAD object models, which are a intensive labour preparation and not always available. Such methods have ability to generalize with novel objects. Motivated by these advantages, our introduced framework computes votes for both objects and hands. Nonetheless, the main point is that the computing process aslo take the interaction conditions between the hand and the held object into account. This helps the model can learn the physical constraints and the interdependences among hands and objects.

In brief, the main contributions of our work are:

- We propose a novel architecture to empower the capability of extracting features from RGB-D images. This network can learn both color and geometric features and then attentionally fuse them together by wisely and selectively magnifying the valued features and weaken the useless one at each pixel.
- We introduce a deep voting-based model to take the strong relationship between hand poses and object shapes into account while computing voting vectors.
- 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 the problem is treat the hand [6, 18, 25, 27, 49] and manipulated object [31, 35, 44, 53] seperately without considering their interdependence. They underestimate the extraordinary relationship between the hand gestures and object shapes. Several approaches overcome this problem by jointly learn the shapes of both hand and object from RGB images. [8] develops two graph convolutional networks for two missions. The first one detects 2D hand joints and 2D object corners, while the second one lifts 2D keypoints to 3D coordinates. [41] proposes attention-guided graph convolution to iteratively share hand and object estimator between two branches for learning the mutual occlusion. [13] looks into photometric consistency between neighboring frames to reconstruct hand-object shape under interactions. [37] handles hand action classification to assist the process of estimating hand-object interactions. [23] 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 3D world is difficultl to accurately proceed due to high degree of non-linearity. In contrary, some methods exploit solely depth images. [5] designs an architecture that firstly predicts hand and object centers and then learn global orientations and grasps of hand configurations while interacting with objects. [12, 52] segments 2D hand and object regions from depth image and then optimizes the reconstruction process of interacting motions. This method, however, learns the depth images by 2D CNN backbone hence cannot radically observe the geometric information. [28] deploys a feedback loop to revise the flawed estimation results using depth images only. RGB-D input data, on the other hand, has received relatively less attention due to how to effectively collaborate two distinctive input format still holds a secret. [21] focuses on the physical laws of hand actions from RGB-D input to benefit hand-object interaction interpretations. [42] 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 [3, 4, 30, 51] and 6D object detection [34, 38, 44] 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 [43]: image layer fusion, feature layer fusion, and output layer fusion. While image layer fusion concates the input data before feeding to CNNs, feature layer fusion means learning color and depth data in two distinguished architecture but sharing the learning process. Output layer fusion, on the other hand, integrate two feature maps which are seperately extracted by two backbone networks. However, fusion RGB-D features for hand-object pose estimation is less attractive because most of mentioned methods have the mutual weak point that is extracting features from depth maps by 2D CNNs. This makes the output 3D spatial feature is latent and unconcious. Motivated by [44], we develop a network of feature layer fusion that can share learning process between two backbones but can export geometric information and geometric contraints by using Pointnet++ [32] backbone for the depth map. In addition, our network adaptively and selectively adjusts features at each pixel before fusing to gain the accurate outcome.

2.3. Voting mechanism for pose estimation

The Hough voting is originally introduced to detect 2D defined shapes [9, 17] and then developed to futher complex computer vision tasks [36, 39]. The deep learning-based voting methods [7, 16, 20, 46] have recently appeared to be a promising approach for object detection and pose estimation due to its robustness and ability to novel object.

3. Method

In Figure 2, we provide an overall pipeline of our method for 3D hand mesh estimation. Our proposed network consists of backbone, Attention, voting and cluster and hand pose Estimation.

3.1. Attentional Fusion

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 [**?**]. In our work, differing from the original work, we conduct PointNet++ [32], 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: Numerous methods integrate color features and geometric features into each other without considering the fact that the distribution of informative features at each position is not equal. Our proposal module termed attentional fusion aims to learn the contributing ability of each pixel feature for effectively fusing procedure. To obtain this, we add learnable weight matrix to either widen or inhibit the pixel feature across the whole image. 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 fusing procedure are used to regress hand joints. Conventional voting methods approaching object pose estimation including hand poses usually vote for the hand center. Whereas, our method computes votes for hand joints points due to the fact that 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 3D object shape. Therefore, such hand joints are neccessary for hand-object interaction learning. We adopt the MANO hand mesh model [33] 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 of 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 3D bouding box are used to represent the object. However, the corner points are actually far away from point on object, leading to the difficulty to infer the physical constrains

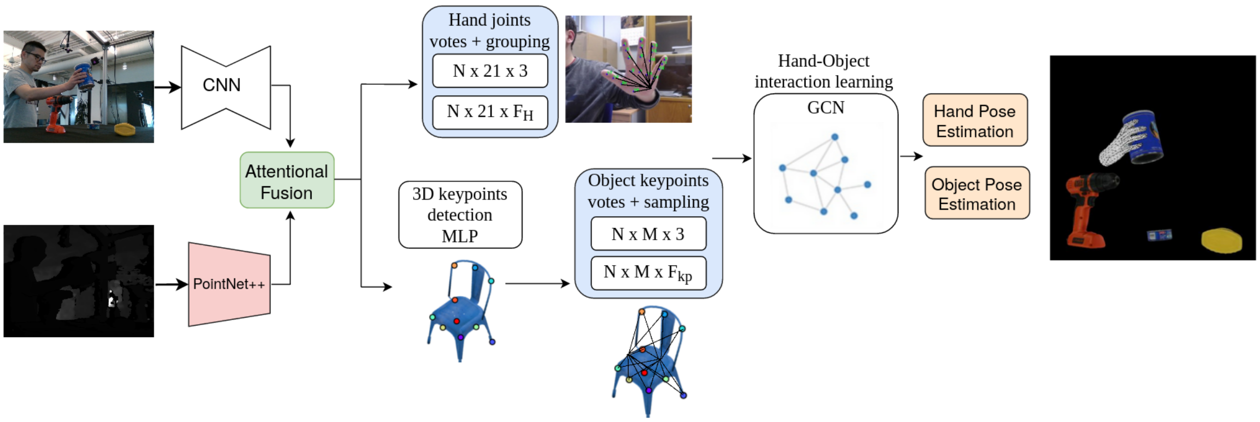

Figure 2: Overview of our proposal. Our method takes both color and depth maps as input data. The color features are extracted by a CNN, while the 3D features are calculated by PointNet++ architecture. These two types of features are then fused together at pixel-level to obtain the new distinctive features. The attention mechanism is applied for the such new features to learn the contribution disparity of the context to the hand pose. The votes are computed and then regressed to estimate the MANO parameters.

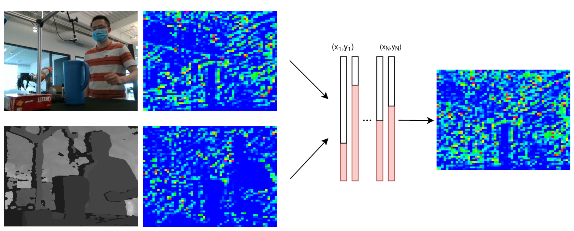

Figure 3: Attentional fusion.

while interacting with the hand. Therefore, we instead select keypoints on the object surfaces that provide ease to learn the hand-object interaction. We use the farthest point sampling (FPS) algorithm to collect the keypoints of objects by initilizing a object mesh center point as the first keypoint and the searching the others by FPS until obtain M keypoints.

Object keypoints voting: In terms of learning the object presence, the attentionally fused features is 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 of 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

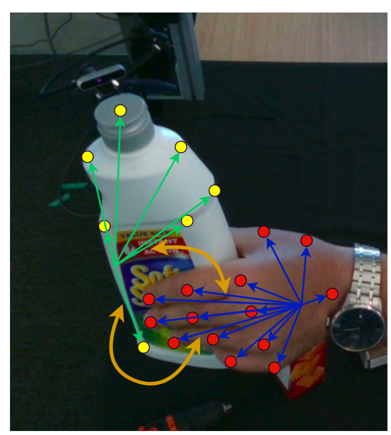

Figure 4: Illustration of the interaction learning between votes for hand keypoints and votes for object keypoints. The red points denote hand keypoints, while the yellow ones denote object keypoints. The blue and green vectors represent hand keypoints and object keypoints votes, respectively.

Hand-object interaction learning: To estimate the hand and object shapes under interactions, voting vectors should be aware of their global neighborhood. Esecially, the object keypoints in a vicinity are intuitively benificial for the predicting hand keypoints and vice cersa. We adopt a graph convolutional network (GCN) for 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 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 adpot the MANO hand mesh model defined 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 fuction 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 differencies 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 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 summarized as below:

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

4. Evaluation

References

[1] 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

[2] 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

[3] 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

[4] 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

[5] 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. 1, 2

[6] 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

[7] Zhipeng Ding, Xu Han, and Marc Niethammer. Votenet: A deep learning label fusion method for multi-atlas segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 202–210. Springer, 2019. 2, 3

[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. 2

[9] RO Duda and PE Hart. ªuse of the hough transform to detect lines and curves in pictures, º comm. *ACM*, 1972. 3

[10] 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

[11] 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

[12] 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. 1, 2

[13] 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. 1, 2

[14] 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. 1

[15] 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. 1

[16] Dinh-Cuong Hoang, Johannes A Stork, and Todor Stoyanov. Voting and attention-based pose relation learning for object pose estimation from 3d point clouds. *IEEE Robotics and Automation Letters*, 7(4):8980–8987, 2022. 2, 3

[17] Paul VC Hough. Machine analysis of bubble chamber pictures. In *Proc. of the International Conference on High Energy Accelerators and Instrumentation, Sept. 1959*, pages 554–556, 1959. 3

[18] 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

[19] 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

[20] Wadim Kehl, Fausto Milletari, Federico Tombari, Slobodan Ilic, and Nassir Navab. Deep learning of local rgb-d patches for 3d object detection and 6d pose estimation. In *European conference on computer vision*, pages 205–220. Springer, 2016. 3

[21] 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

[22] 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

[23] 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. 1, 2

[24] Yao Lu and Walterio W Mayol-Cuevas. Understanding egocentric hand-object interactions from hand pose estimation. *arXiv preprint arXiv:2109.14657*, 2021. 1

[25] 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

[26] 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

[27] 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

[28] 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. 1, 2

[29] 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. 1

[30] 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

[31] 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

[32] 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. 2, 3

[33] Javier Romero, Dimitrios Tzionas, and Michael J Black. Embodied hands: Modeling and capturing hands and bodies together. *arXiv preprint arXiv:2201.02610*, 2022. 3

[34] 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

[35] 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

[36] Teresa M Silberberg, Larry Davis, and David Harwood. An iterative hough procedure for three-dimensional object recognition. *Pattern Recognition*, 17(6):621–629, 1984. 3

[37] 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. 1, 2

[38] 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

[39] Federico Tombari and Luigi Di Stefano. Object recognition in 3d scenes with occlusions and clutter by hough voting.

In *2010 Fourth Pacific-Rim Symposium on Image and Video Technology*, pages 349–355. IEEE, 2010. 3

[40] 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

[41] 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. 1, 2

[42] 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

[43] 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

[44] 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

[45] 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. 1

[46] Yangzheng Wu, Mohsen Zand, Ali Etemad, and Michael Greenspan. Vote from the center: 6 dof pose estimation in rgb-d images by radial keypoint voting. *arXiv preprint arXiv:2104.02527*, 2021. 2, 3

[47] 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

[48] 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

[49] 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

[50] 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

[51] 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

[52] 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. 1, 2

[53] 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

[54] 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