Coarse-to-Fine Hand–Object Pose Estimation with Interaction-Aware Graph Convolutional Network

Maomao Zhang, Ao Li *, Honglei Liu and Minghui Wang

School of Information Science and Technology, University of Science and Technology of China, Hefei 230027, China; zmm1@mail.ustc.edu.cn (M.Z.); lhl0796@mail.ustc.edu.cn (H.L.); mhwang@ustc.edu.cn (M.W.)
* Correspondence: aoli@ustc.edu.cn

Abstract: The analysis of hand–object poses from RGB images is important for understanding and imitating human behavior and acts as a key factor in various applications. In this paper, we propose a novel coarse-to-fine two-stage framework for hand–object pose estimation, which explicitly models hand–object relations in 3D pose refinement rather than in the process of converting 2D poses to 3D poses. Specifically, in the coarse stage, 2D heatmaps of hand and object keypoints are obtained from RGB image and subsequently fed into pose regressor to derive coarse 3D poses. As for the fine stage, an interaction-aware graph convolutional network called InterGCN is introduced to perform pose refinement by fully leveraging the hand–object relations in 3D context. One major challenge in 3D pose refinement lies in the fact that relations between hand and object change dynamically according to different HOI scenarios. In response to this issue, we leverage both general and interaction-specific relation graphs to significantly enhance the capacity of the network to cover variations of HOI scenarios for successful 3D pose refinement. Extensive experiments demonstrate state-of-the-art performance of our approach on benchmark hand–object datasets.

Keywords: hand–object pose estimation; deep learning; graph convolutional network; coarse-to-fine

1. Introduction

Hand–object interaction (HOI) is a primary part of human daily behavior. Therefore, analyzing hand–object interaction and identifying the poses of hand and object are crucial for understanding and imitating human daily behavior. Recently, hand–object pose estimation has attracted considerable attention in a variety of applications, for example, augmented and virtual reality, robotics, and human–computer interaction.

Although estimating hand pose [1–7] and object pose [8–11] in isolation has made remarkable success, jointly estimating both hand and manipulated object poses from a single RGB image remains a challenging task. Particularly, complicated HOI scenarios bring in various issues including not only complex pose variations and self-occlusions that commonly occur in hand-only or object-only pose estimation, but also severe mutual occlusion between hand and manipulated object [12]. Meanwhile, it should also be pointed out that considering the high relevance of hand and manipulated object is helpful for addressing the above issues, since object shape usually enforces physical constraints on hand grasps while hand pose could give cues to object pose and category. Accordingly, there is an urgent need for improving performance in hand–object pose estimation by fully modeling the kinematic relations between hand and object poses in HOI.

In early works [13–15], the relations between hand and manipulated object are not taken into account, given the fact that these approaches estimate the poses of hand and object separately. Recently, some works [12,16–19] start to employ the relations between hand and object as evidence for joint hand–object pose estimation. Despite great progress, most existing approaches try to model hand–object relations during converting 2D poses to 3D poses. However, it is worth noting that generating 3D poses from 2D coordinates is
inherently ill-posed problem [20]. In particular, there may exist multiple valid 3D interpretations corresponding to the same 2D representation. In consequence, these methods have limitations in precisely catching and leveraging the relations between hand and object to estimate 3D hand–object poses. Differently, in this work we focus on modeling hand–object relations effectively in the process of refining 3D poses.

To address the above limitations, we propose a novel coarse-to-fine framework for joint hand–object pose estimation, which can progressively improve the performance via multi-step reasoning. In particular, given a monocular RGB image containing hand and object, we first estimate heatmaps of hand joints and object corners in the 2D image, which are then passed through pose regressor to generate coarse 3D poses of hand joints and object corners. After that, an effective graph-based model named Interaction-aware Graph Convolutional Network (InterGCN), is proposed to perform pose refinement by fully leveraging the hand–object relations in 3D context. In this manner, the coarse-to-fine cascade enables our approach to increase the robustness of hand–object pose estimation and achieve more accurate 3D pose results.

Instead of catching hand–object relations during converting 2D poses to 3D poses, we show that in the process of refining coarse 3D poses, the proposed InterGCN can significantly improve performance by explicitly modeling the relations between hand and object. To model rich relations of hand–object interaction, we construct a graph where the nodes are hand joints and object corners. Based on the prior that keypoints have strong relations when they are close to each other, we construct a graph by nearest neighbor analysis on the training dataset, which reflects a general pattern across HOI scenarios. However, the relations between hand and object often change dynamically according to different HOI scenarios, which means that the general graph is insufficient to model complex hand–object relations. To relieve this problem, we propose to learn an additional interaction-specific graph from representations of graph nodes, which delineates a unique pattern of hand–object relations for each HOI scenario. In an effective and specialized way, this interaction-specific graph can capture variations of HOI scenarios for successful 3D pose refinement.

The core contributions of our work are summarized as follows:

- We propose a novel deep learning framework for hand–object pose estimation, which can progressively improve the model’s performance using a coarse-to-fine strategy;
- We introduce Interaction-aware Graph Convolutional Network to explicitly model rich and dynamic hand–object relations to optimize the coarse pose results;
- Extensive experimental results on benchmarks demonstrate that our approach outperforms state-of-the-art methods.

2. Related Works

Our work closely relates to joint 3D hand–object pose estimation and graph convolutional networks for dealing with graph-structured data.

2.1. Hand–Object Pose Estimation

Early approaches [13–15] focus on estimating the pose of hand or object in isolation, which nevertheless may ignore strong relations between hand and manipulated object. Subsequently, by utilizing interaction of hand and object as additional constraints, several works [17,21,22] try to jointly estimate hand and object poses with multi-view RGB or depth input. However, due to various limitations, such as the high cost for a multiple RGB sensors system and the huge power consumption using active depth sensors [12], some researchers shift their focus to RGB-based hand–object pose estimation methods. Several recent works [16,23] employ generative methods, e.g., the MANO [24], to extract geometrical dynamic constraints of hand and object. Nevertheless, such methods usually need extra dense annotations that are usually difficult to obtain in practice.

Given RGB images, there have been some other effective approaches [12,18,19] for hand–object poses estimation using relations between hand joints and object corners.
Tekin et al. [18] adopt a unified 3D detection framework to directly output the poses of hand and object without explicitly considering the physical constraints on hand–object interaction. More recently, some methods try to model the hand–object relations during the process of converting 2D poses to 3D. Doosti et al. [19] first regress 2D keypoint locations of hand and object, and then adopt an adaptive graph convolutional network to learn a non-linear mapping between 2D and 3D poses. Instead of using predefined adjacency matrix in graph convolution layer, they use a parameterized matrix learned from training data to capture relations between hand and object. Huang et al. [12] propose a non-autoregressive transformer module to lift 2D poses to 3D where multi-head self-attention layers are employed to model correlations between hand joints and object corners. However, generating 3D poses from 2D coordinates is inherently ill-posed [20] since there may exist multiple valid 3D interpretations for a single 2D representation, which in consequence may cause unreasonable hand–object relations and invalid results. Instead of obtaining relations during the process of lifting 2D poses to 3D, we propose InterGCN to model hand–object relations and leverage them to refine 3D poses.

2.2. Graph Convolutional Networks

Graph convolutional networks (GCNs) are introduced to capture inherent dependencies between nodes of graph-structured data. In general, the principle of constructing GCN follows two mainstreams: spectral perspective [25,26] and spatial perspective [27–29]. Spectral GCNs use the spectral representations of graphs to perform graph convolution operations. Cai et al. [20] adopt a spectral GCN to handle graphs with predefined and fixed topology for 3D pose estimation, which nevertheless lacks the flexibility and capacity to model latent relations among nodes. In comparison, spatial GCNs directly perform convolution filters on graph vertexes and their neighbors and usually can handle topology-varied graphs. For hand–object pose estimation, Doosti et al. [19] utilize an adaptive GCN to model subtle relationships between joints by constructing a parameterized relation graph. However, this approach tries to learn a general graph for different HOI scenarios, which may be suboptimal for all the samples due to the complexity and variability of HOI in real-world applications. Instead, we propose InterGCN that takes advantages of not only prior knowledge of general hand–object relations, but also interaction-specific relations automatically learned from representations of hand joints and objects corners.

3. Methods

Given a single RGB image containing hand and object, our goal is to figure out 3D hand–object poses, where hand pose is defined by a set of joint coordinates \( p_h = (x_m, y_m, z_m)_{m=1}^M \) in 3D space while object pose is represented by a set of bounding box corner coordinates \( p_o = (x_n, y_n, z_n)_{n=1}^N \). \( M \) and \( N \) are the number of hand joints and object corners, respectively. In our case, we choose \( M = 21 \) and \( N = 8 \) since skeletal model of the hand used in this work has 21 joints and object pose is represented by 8 object bounding box corners. The hand model with 21 joints is shown in Figure 1. The overall scheme of the proposed coarse-to-fine framework is presented in Figure 2. Our method consists of two stages: coarse stage and fine stage. In the coarse stage, image is passed through a 2D pose estimation network to obtain the heatmap of each keypoint, which is used as the input of pose regressor to generate the coarse 3D poses of hand and object. As for the fine stage, InterGCN is proposed to leverage both general hand–object relations and interaction-specific relations learned from representations of hand joints and object corners to refine the predicted coarse poses in 3D context. Details are described in the later sections.

3.1. Heatmap-Based Coarse Pose Generation

In this stage, we formulate localization of 2D keypoints as estimation of 2D heatmaps. The 2D heatmap \( H_k \in \mathbb{R}^{H \times W} \) represents the likelihood of the \( k \)th keypoint at each pixel location, where \( H, W \) denote the height and width of heatmap, respectively. We employ a 2D heatmap estimation network with encoder–decoder architecture based on the Con-
volutional Pose Machines (CPMs) by Wei et al. [30]. Given image feature representations generated by the encoder, an initial heatmap is predicted and iteratively enhanced in resolution. The number of total iterations are fixed at 3. Furthermore, inspired by [16], we utilize a two-branch network on top of a common CNN encoder to estimate 2D poses of hand and object separately. Note that the two branch networks are different and have independent learnable parameters. Subsequently, a simple yet effective pose regressor similar to [3] is adopted to generate coarse 3D poses of hand and object from the obtained 2D heatmaps. In addition to the generated 2D heatmaps, the intermediate image features in 2D pose estimation network are also fed into the pose regressor by following [31].

Figure 1. The illustration of hand model with 21 joints.

Figure 2. The overview of the proposed coarse-to-fine framework. Our method consists of two stages: coarse stage and fine stage. In the coarse stage, hand and object keypoint heatmaps are first obtained from input image. Subsequently, the heatmaps are passed to the pose regressor to generate coarse 3D hand–object poses. As for the fine stage, InterGCN is introduced to refine coarse hand–object poses estimated in the previous stage by leveraging the general relation graph (Section 3.2.2) and interaction-specific relation graph between hand joints and object corners in 3D context.

3.2. Interaction-Aware Graph Convolutional Network
3.2.1. Graph Convolution

In this section, we introduce graph convolution in InterGCN, accounting for the heterogeneous types of nodes.
We first describe a typical Graph Convolutional Network introduced by Kipf and Welling [25]. Let $G = (V, E)$ donate a graph where $V$ is a set of $|V|$ nodes and $E$ represents the set of edges between the nodes in $V$. Graph convolutional operation can be formulated as:

$$X^{(l+1)} = \sigma(\tilde{A}X^{(l)}W^{(l)}), \tilde{A} = \tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}$$ (1)

where $X^{(l)} \in \mathbb{R}^{|V| \times d^{(l)}}$ and $X^{(l+1)} \in \mathbb{R}^{|V| \times d^{(l+1)}}$ are input and output node representations at $l$th layer with dimensionality of $d^{(l)}$ and $d^{(l+1)}$, respectively. $X^{(l)}$ is a $|V| \times d^{(l)}$ input feature matrix. Each row in $X^{(l)}$ is a $d^{(l)}$-dimensional vector representing a graph node. $W^{(l)} \in \mathbb{R}^{d^{(l)} \times d^{(l+1)}}$ is the learnable weight matrix, $\sigma$ is activation function (e.g., ReLU). $\tilde{A} = A + I$ is the new self-loop adjacency matrix where $A \in \mathbb{R}^{|V| \times |V|}$ is the adjacency matrix of $G$ and $I$ is identity matrix. $\tilde{D}$ is the diagonal node degree matrix of $\tilde{A}$. $\tilde{A}$ is the symmetric normalized version of $A$. The $A$ is a $|V| \times |V|$ matrix and represents relations among graph nodes, i.e., the value of $\tilde{A}_{ij}$ represents relation between node $i$ and node $j$.

The edge information in $\tilde{A}$ is used to aggregate features in each layer of GCN.

Despite its effectiveness, the direct application of GCN in hand–object pose estimation may limit capability of the network due to node heterogeneity issue. Specifically, there are two different types of nodes in HOI scenarios: hand joints $h$ and object corners $o$, which are associated with distinctive pose characteristics as well as feature spaces. To better adapt GCN for hand–object pose estimation, we exploit heterogeneous graph convolution, which takes into account the difference of node types and adopts respective weight matrices for feature transformation. The operation can be represented as:

$$X^{(l+1)}_{\tau} = \sigma(\sum_{\tau' \in \Gamma} A_{\tau\tau'}X^{(l)}_{\tau'}W^{(l)}_{\tau'})$$ (2)

where $\tau, \tau' \in \Gamma = \{h, o\}$, $A_{\tau\tau'} \in \mathbb{R}^{|V_{\tau}| \times |V_{\tau'}|}$ is the submatrix of $A$, with rows and columns denoting the nodes with type $\tau$ and their neighboring nodes with type $\tau'$, respectively. The node features $X^{(l+1)}_{\tau'}$ are generated by aggregating features of their neighboring nodes $X^{(l)}_{\tau'}$ with different types $\tau'$ using different weight matrix $W^{(l)}_{\tau'} \in \mathbb{R}^{d^{(l)} \times d^{(l+1)}}$. The weight matrix $W^{(l)}_{\tau'}$ takes into account difference of distinct feature spaces and maps them into an implicit common space.

As the core of graph convolution, the adjacency matrix is utilized in InterGCN to explicitly model rich relations of hand and object. Specifically, it consists of two main parts: $A^h$ representing general relation graph and $A^o$ representing interaction-specific relation graph. In the following sections, we describe how to construct these two graphs and leverage them to update the node representations in detail.

### 3.2.2. General Relation Graph

To explicitly model hand–object relations, the general relation graph $A^G$ is disentangled into non-overlapped subgraphs, i.e., hand-to-hand subgraph $A^G_{hh}$, object-to-object subgraph $A^G_{oo}$, object-to-hand subgraph $A^G_{ho}$ and hand-to-object subgraph $A^G_{oh}$. Specifically, the normalized adjacency matrix $A^G_{hh} \in \mathbb{R}^{M \times M}$ and $A^G_{oo} \in \mathbb{R}^{N \times N}$ indicate the physical structure of hand skeleton and object bounding box, respectively. Meanwhile, we adopt a data-driven way (Algorithm 1) to build $A^G_{ho}, A^G_{oh}$, based on the prior that keypoints have strong relations when they are close to each other. Specifically, we assemble 3D hand joint coordinates and object corner coordinates of the training dataset into data tensor $P^h \in \mathbb{R}^{F \times M \times 3}$ and $P^o \in \mathbb{R}^{F \times N \times 3}$, respectively, where $F$ is the number of samples and the last dimension represents the $(x, y, z)$ coordinates. For hand joint $i$ and $j$, if $j$ is the closest object corner to it, then corner $j$ is considered as the one most closely related to joint $i$ and $A^G_{ho}^{F(i,j)} = 1$. Subsequently, we normalize matrix $A^G_{ho}$ by scaling the rows so that the row sums are all equal to one. With these operations, we encode general object-to-hand relations in $A^G_{ho}$. 
That is, $A^{s,(ij)}_{ho}$ represents the relative importance of the object corner $j$ to the hand joint $i$. In the same way, we can obtain subgraph $A^{o}_{oh}$ that reflects general hand-to-object relations.

### Algorithm 1 Construct relation subgraphs $A^{s}_{ho}$, $A^{o}_{oh}$

**Input**: 3D hand pose labels $P^h \in \mathbb{R}^{F \times M \times 3}$, 3D object pose labels $P^o \in \mathbb{R}^{F \times N \times 3}$, $F$ is sample number of training dataset.

**Output**: relation matrix $A^{s}_{ho} \in \mathbb{R}^{M \times N}$, $A^{o}_{oh} \in \mathbb{R}^{N \times M}$.

1. Initialize $A^{s}_{ho}$, $A^{o}_{oh}$ with zeros.
2. for each pose $(p^h, p^o) \in (P^h, P^o)$ do
   3. for $i = 0$ to $M$ do
      4. $j^* = \arg\min_j \|p^h_i - p^o_j\|_2$
      5. $A^{s,(ij^*)}_{ho} = 1$
   6. end for
   7. for $i = 0$ to $N$ do
      8. $j^* = \arg\min_j \|p^o_i - p^h_j\|_2$
      9. $A^{o,(ij^*)}_{oh} = 1$
  10. end for
  11. end for

#### 3.2.3. Interaction-Specific Relation Graph

Graph $A^s$ implies the relations between the hand joints and the object corners. Nevertheless, it only accounts for a general pattern across HOI scenarios, which may not be optimal for all the samples. For example, in HOI scenario “closing juice bottle”, the relations between fingertips and bottle cap should be stronger, but it is not true for some other scenarios, such as “pouring juice bottle” and “putting salt”. This fact suggests that hand-object relations should be interaction-specific and data dependent, which unfortunately is not supported in graph $A^s$.

To solve this issue, we further present an interaction-specific relation graph $A^s$ which is unique for each HOI scenario with hand- and object-centric attention mechanism. As shown in Figure 3, given representations of hand joints and object corners $X_h \in \mathbb{R}^{M \times d}$, $X_o \in \mathbb{R}^{N \times d}$, three linear layers are applied to generate query vector $Q_h \in \mathbb{R}^{M \times d}$ for hand joints, and two key vectors: $K_h \in \mathbb{R}^{M \times d}$ and $K_o \in \mathbb{R}^{N \times d}$ for hand joints and object corners, respectively. They are then multiplied to obtain a $M \times M$ hand-to-hand subgraph $A^{s}_{hh}$ and a $M \times N$ object-to-hand subgraph $A^{s}_{ho}$. $A^{s,(ij)}_{hh}$ and $A^{s,(ij)}_{ho}$ indicate the impact of node $j$ with type $h$ and node $j$ with type $o$ to node $i$ with type $h$, respectively. The value of subgraphs is normalized to 0–1, which is used as the soft edge of two nodes. With a softmax operation, we can calculate $A^{s}_{hh}$ and $A^{s}_{ho}$ as follows:

$$
A^{s,(ij)}_{hh} = \frac{\exp(X^j_h W^Q_h (X^j_h W^K_h)^T)}{\sum_{j=1}^{M} \exp(X^j_h W^Q_h (X^j_h W^K_h)^T)}
$$

$$
A^{s,(ij)}_{ho} = \frac{\exp(X^j_h W^Q_h (X^j_o W^K_o)^T)}{\sum_{j=1}^{N} \exp(X^j_h W^Q_h (X^j_o W^K_o)^T)}
$$

where $W^Q_h$, $W^K_h$ and $W^K_o$ are the parameters of the linear layers for computing the hand query $Q_h$, hand key $K_h$ and object key $K_o$. $X^j_h$, $X^j_o$ and $X^j_i$ are node features. In the same manner, we can construct object-to-object subgraph $A^{o}_{oo}$ and hand-to-object subgraph $A^{o}_{oh}$ with object-centric attention mechanism.
3.2.4. Feature Aggregation

With the general and the interaction-specific relation graphs, we can update the representations of hand joints and object corners by aggregating features of other graph nodes. As shown in Figure 3, considering the distinct node types in InterGCN, two parallel linear layers are applied to encode features of hand joints and object corners, respectively. After that, we aggregate features from not only hand joints but also object corners by leveraging both general relation graph and interaction-specific relation graph. This process is computed by replacing Equation (2) with the following layer-wise propagation rule:

\[
X_h^{(l+1)} = \sigma((A_{hh}^g + A_{hh}^s)X_h^{(l)}W_h^i + (A_{ho}^g + A_{ho}^s)X_o^{(l)}W_o^i)
\] (4)

Similarly, we can update the representations of object corners as:

\[
X_o^{(l+1)} = \sigma((A_{oo}^g + A_{oo}^s)X_o^{(l)}W_o^i + (A_{oh}^g + A_{oh}^s)X_h^{(l)}W_h^i)
\] (5)

3.3. Loss Functions

We train our model with three loss functions: heatmap loss \(L_{hm}\), coarse 3D pose loss \(L_c\), and refined 3D pose loss \(L_f\).

3.3.1. Heatmap Loss

\[L_{hm} = \sum_{j=1}^{M+N} \|H_j - \hat{H}_j\|_2^2\]

where \(H_j\) and \(\hat{H}_j\) denote ground-truth and estimated heatmap, respectively. We set the heatmap resolution as \(32 \times 32\) pixels. The ground-truth heatmap is defined as 2D Gaussian with standard deviation of 1 pixel centered on the ground-truth 2D location.

3.3.2. 3D Pose Loss

Three-dimensional pose loss composes of two parts: coarse pose loss \(L_c = \sum_{j=1}^{M+N} \|P_j - \hat{P}_j^c\|_2^2\) and refined pose loss \(L_f = \sum_{j=1}^{M+N} \|P_j - \hat{P}_j^f\|_2^2\), where \(P_j, \hat{P}_j^c\) and \(\hat{P}_j^f\) denote ground-truth, coarse pose and refined pose with InterGCN, respectively.

The total loss function for training is:

\[
\mathcal{L} = L_{hm} + \alpha L_c + \beta L_f
\] (6)
where \( \alpha = 1 \), and \( \beta = 2 \) are weight coefficients to balance different loss functions.

4. Experiments

4.1. Datasets

First-Person Hand Action (FPHA) Dataset [32] is a large-scale and commonly used dataset for 3D hand–object pose estimation, containing 1175 videos of hand actions performed by 6 subjects from egocentric point-of-view. Three-dimensional hand poses are obtained automatically with visible magnetic sensors. A subset of FPHA contains object pose labels for 4 objects (juice bottle, liquid soap, milk, and salt) involving a variety of action categories, which is denoted as FPHA-HO and adopted here for performance evaluation. Briefly, position and orientation with 6 degrees of freedom are obtained using six magnetic sensors attached to subject’s hand. Then, hand pose is derived from inverse kinematics of a defined hand model. As for object pose, one more sensor is attached to the closest point to the center of mass.

HO-3D [33] is a recently released hand–object dataset, which consists of video sequences of hands interacting with objects from third-person point-of-view. HO-3D is a markerless hand–object dataset containing 77 k annotated frames, 65 sequences, 10 subjects, and 10 objects. Then, 66 k frames with 3D hand–object pose labels are used as training set. Note that hands are only annotated with the wrist coordinates in testing set containing 11 k frames.

ObMan [16] is a synthetic dataset containing 141 k training images, 6 k validation images, and 6 k testing images. Images in this dataset are obtained by rendering 3D hand meshes with 8 objects from ShapeNet [34]. We pretrain the proposed model on this synthetic dataset, and then fine-tune on above two datasets.

4.2. Evaluation Metrics

Following [12], mean error in Euclidean space between estimated joint coordinates and ground truth is used to evaluate 3D hand poses. The percentage of correct keypoints (3D PCK) is also adopted to measure accuracy of hand pose estimation. For evaluating object pose, we utilize the percentage of correct poses (PCP). In this metric, if the 2D projection error of object corners is less than a given threshold, object pose is considered to be estimated correctly. Moreover, we also provide the area under the curve (AUC) on the PCK for hand pose and the PCP for object pose. Euclidean distance is used to calculate 2D and 3D error. We use mean error to determine PCP and PCK.

4.3. Implementation Details

Using a resized 256 × 256 HOI image as input, the 2D pose estimation network generates keypoint heatmaps with resolution 32 × 32. We adopt ResNet-18 [35] as the encoder of 2D pose estimation network. Instead of training the whole network together, we first train the 2D pose estimation network and the pose regressor for 60 epochs with an initial learning rate of 0.01 and multiply it by 0.1 every 20 epochs. Then, we optimize InterGCN for another 60 epochs, starting from a learning rate of 0.01 with a shrink factor of 0.1 every 20 epochs. Finally, we fine-tune the whole network for 40 epochs, and the learning rate starts from 0.001 and is multiplied by 0.1 every 20 epochs. We use Adam as network optimizer and set the batch size to 64. The hyperparameters are selected based on grid search strategy over validation dataset. Specifically, the initial learning rate is selected from \( \{0.1, 0.01, 0.001, 0.0001\} \) and batch size is selected from \( \{16, 32, 64, 128\} \). All experiments are conducted on GeForce GTX 1080Ti GPU using PyTorch framework. For both FPHA-HO [32] and HO-3D [33] datasets, we use official train and test splits, and pretrain on ObMan [16].

4.4. Comparisons with the State-of-the-Arts

In this section, we report the performance of our approach and compare with the state-of-the-art methods. Similar to [12], the absolute 3D poses in camera coordinate system
(c.s.) are denoted as Abs. while the root-relative 3D poses are represented as Rel. Hand pose and object pose are denoted as HP and OP, respectively.

We first compare the proposed method with others on FPHA-HO [32] using the mean 3D distance error of hand and object. As shown in Table 1, our approach compares favorably with the state-of-the-art approaches H+O [18] and HOT-Net [12], especially for the relative hand pose (8.32 vs. 10.41). In terms of the PCP metric of object pose, as illustrated in Figure 4, our approach outperforms H+O and HOT-Net by a large margin over all the thresholds. For 3D PCK curve of hand pose, we compare with H+O and HOT-Net as well as another depth-based method [32]. The results in Figure 4 demonstrate the promising performance of the proposed method.

![Figure 4. Comparison with state-of-the-art methods on PCP curve and 3D PCK curve.](image)

### Table 1. Comparison with state-of-the-art methods on FPHA-HO. The mean 3D distance error (mm) is used as metric (Lower is better).

| Model       | Abs. HP Error | Rel. HP Error | Abs. OP Error |
|-------------|---------------|---------------|---------------|
| H+O         | 15.81         | -             | 24.89         |
| HOT-Net     | 15.18         | 10.41         | 21.37         |
| Ours Coarse | 16.25         | 10.04         | 27.31         |
| Ours        | 14.97         | 8.32          | 23.07         |

The Area Under Curve (AUC) scores on the PCK curve for hand pose and the PCP curve for object pose are also reported on FPHA-HO [32] and HO-3D [33]. Since hands in the testing dataset are only labeled with wrist coordinates (without complete hand pose annotations), the AUC on PCK for HO-3D are just for wrist keypoint. As seen in Table 2, the AUC score of our method reaches 0.839 for hand pose and 0.654 for object pose on FPHA-HO. Compared to HOT-Net [12], the proposed method achieves 1.21% and 9.91% gains, which indicates the effectiveness of our method. Even on less constrained and more complex dataset HO-3D, our light weight model remains competitive against HOT-Net [12].

### 4.5. Ablation Study

In this section, FPHA-HO [32] is used to identify the effects of the components of our method.
Table 2. Comparison with state-of-the-art methods on both FPHA-HO and HO-3D using the AUC scores (Higher is better) on 3D PCK curve and PCP curve.

| Dataset | Model  | AUC on PCK | AUC on PCP |
|---------|--------|------------|------------|
| FPHA-HO | HOT-Net | 0.829      | 0.595      |
|         | Ours   | 0.839      | 0.654      |
| HO-3D   | HOT-Net | 0.819      | 0.567      |
|         | Ours   | 0.805      | 0.583      |

4.5.1. Effectiveness of Coarse-to-Fine Strategy

This self-comparison experiment is carried out to show the significant improvement by adopting the coarse-to-fine strategy in hand–object pose estimation. “Ours Coarse” means the coarse 3D pose results without using InterGCN for refinement. As shown in Table 1, compared to “Ours Coarse”, our full coarse-to-fine model remarkably reduces the mean 3D distance error by 1.28 mm (from 16.25 to 14.97), 1.72 mm (from 10.04 to 8.32), and 4.24 mm (from 27.31 to 23.07) for Abs. HP, Rel. HP, and Abs. OP, respectively. Obviously, the results of this contrast experiment prove that coarse-to-fine cascade can significantly boost the performance of hand–object pose estimation.

4.5.2. Effectiveness of Relation Graphs

As introduced in the method section, there are two types of graphs representing hand–object relations in each layer of InterGCN, i.e., general relation graph $A^g$ and interaction-specific relation graph $A^s$. To examine the effect of using different relation graphs for hand–object pose estimation, two sets of experiments are conducted. As presented in Table 3, the results show that the two relation graphs are both beneficial for hand–object pose estimation and that deleting any one of the two graphs will degrade the performance. It is also worth noting that removing interaction-specific relation graph leads to a significantly greater decline in performance, demonstrating that modeling unique relations for each HOI scenario plays a crucial role in hand–object pose estimation. With all relation graphs added together, the model achieves the best performance. In addition, Figure 5 breaks out the distance error of each hand joint, which is in line with the above results and further illustrates the effectiveness of relation graphs.

Table 3. Comparison of mean 3D distance error (mm) when using InterGCN without general relation graph $A^g$ or interaction-specific relation graph $A^s$ on FPHA-HO.

| Method            | Abs. HP Error | Rel. HP Error | Abs. OP Error |
|-------------------|---------------|---------------|---------------|
| Ours Coarse       | 16.25         | 10.04         | 27.31         |
| Ours w/o $A^g$    | 15.59         | 9.14          | 24.83         |
| Ours w/o $A^s$    | 15.26         | 8.79          | 23.89         |
| Ours              | 14.97         | 8.32          | 23.07         |

4.5.3. Visualization of Relation Graphs

Figure 6 displays the general relation graph in the FPHA-HO dataset and two examples of the interaction-specific relation graph learned by InterGCN in two different HOI scenarios: “pouring juice bottle”, and “closing juice bottle”. The gray scale of each element in the matrix represents the strength of the relation between two nodes. From the right matrix, we can see from the columns that some object corners (node 21, 23) have stronger effect on the fingertips (e.g., node 4, 8, 12) in “closing juice bottle”. Such relations does not seem to exist in “pouring juice bottle”, as illustrated in the middle matrix. This difference shows that the learned interaction-specific graph can explicitly model unique relations between hand and object for different HOI cases.
Figure 5. Three-dimensional distance error per hand joints. Note that M, P, and D denote the 3 consecutive joints located between the wrist and fingertip (T) on each finger, in their order.

Figure 6. Visualization of the general relation graph and the learned interaction-specific relation graphs in different HOI scenarios. The left matrix is the general relation graph in the FPHA-HO dataset. The middle matrix is an example of the interaction-specific relation graph learned by our InterGCN in “pouring juice bottle”. The right matrix is another example of the interaction-specific relation graph for “closing juice bottle”.

4.6. Qualitative Results

Some qualitative results of our method on the FPHA-HO are illustrated in Figure 7, which also show the effectiveness of coarse-to-fine framework. It is clearly that more reliable results can be consistently achieved in the fine stage, despite of the complexity and versatility of hand and object poses. For example, from the visualized coarse 3D pose results in the first row, it can be seen that the thumb and index fingers are pinched together, which is incorrect for the HOI scenario “closing juice bottle”. In comparison, the visualized refined 3D poses suggest that the thumb and index fingers correctly make a “C” shape to grasp and tighten the juice bottle cap, which is remarkably consistent with the ground truth of hand poses. Taken together, these results further suggest that our method has a strong capability in estimating hand–object poses for different HOI scenarios.
Figure 6. Qualitative results of our method on the FPHA-HO dataset. From left to right: input image, coarse pose results without using InterGCN, refined pose results, and ground truth. The proposed coarse-to-fine framework can handle diverse HOI scenarios with different objects and severe occlusions.

Figure 7. Qualitative results of our method on the FPHA-HO dataset. From left to right: input image, coarse pose results without using InterGCN, refined pose results, and ground truth. The proposed coarse-to-fine framework can handle diverse HOI scenarios with different objects and severe occlusions.

4.7. Runtime

In this section, we report training and inference time of our model. Our model is easy to train. It takes about 18 and 40 h to train the proposed coarse-to-fine framework on FPHA-HO and HO-3D using a GeForce GTX 1080Ti GPU, respectively. As for inference time, it takes about 0.02 s to estimate hand–object poses from a single sample image.

5. Conclusions

In this paper, a novel deep-learning framework is proposed to jointly estimate hand–object poses, which can progressively improve the performance with coarse-to-fine strategy. We focus on refining the coarse poses with InterGCN by modeling and leveraging the rich and dynamic relations between the hand and the manipulated object. In addition to
a general relation graph across HOI scenarios, we also propose to utilize an interaction-specific relation graph to catch unique pattern for each interaction sample. With this manner, our method can effectively increase the capacity of the network to cover variations of HOI scenarios for effective 3D pose refinement. We carry out extensive experiments and detailed ablation studies to show the superiority of our method. In the future, we will try to evaluate person-independent hand pose estimation with customized training and testing data split.

Author Contributions: Conceptualization, M.Z. and A.L.; methodology, M.Z. and A.L.; software, M.Z. and H.L.; validation, M.Z. and H.L.; writing—original draft preparation, M.Z.; writing—review and editing, A.L. and M.W.; supervision, A.L.; funding acquisition, A.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work was partly supported by Scientific Research Foundation for the Talents of USTC and the National Natural Science Foundation of China (61971393, 61871361).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data available in a publicly accessible repository. Data citation: [dataset] Garcia-Hernando, G.; Yuan, S.; Baek, S.; Kim, T.-K. 2018. First-Person Hand Action (FPHA) Dataset; https://guiggh.github.io/publications/first-person-hands (accessed on 2 July 2020); DOI: 10.1109/CVPR.2018.00050; [dataset] Hampali, S.; Rad, M.; Oberweger, M.; Lepetit, V. 2020. HO-3D; https://www.tugraz.at/index.php?id=40231 (accessed on 26 August 2020); Version 2; DOI: 10.1109/CVPR42600.2020.00326; [dataset] Hasson, Y.; Varol, G.; Tzionas, D.; Kalevatykh, I.; Black, M.J.; Laptev, I.; Schmid, C. 2019. ObMan; https://hassony2.github.io/obman (accessed on 22 July 2020); DOI: 10.1109/CVPR.2019.01208.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Tompson, J.; Stein, M.; Lecun, Y.; Perlin, K. Real-time continuous pose recovery of human hands using convolutional networks. *ACM Trans. Graph.* 2014, 33, 1–10. [CrossRef]

2. Oberweger, M.; Wohlhart, P.; Lepetit, V. Training a Feedback Loop for Hand Pose Estimation. In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015.

3. Zimmermann, C.; Brox, T. Learning to Estimate 3D Hand Pose from Single RGB Images. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017.

4. Iqbal, U.; Molchanov, P.; Gall, T.B.J.; Kautz, J. Hand Pose Estimation via Latent 2.5D Heatmap Regression. In Proceedings of the 2018 European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018.

5. Wan, C.; Probst, T.; Van Gool, L.; Yao, A. Dense 3D Regression for Hand Pose Estimation. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018.

6. Ge, L.; Ren, Z.; Li, Y.; Xue, Z.; Wang, Y.; Cai, J.; Yuan, J. 3D Hand Shape and Pose Estimation from a Single RGB Image. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019.

7. Huang, W.; Ren, P.; Wang, J.; Qi, Q.; Sun, H. AWR: Adaptive Weighting Regression for 3D Hand Pose Estimation. In Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), New York, NY, USA, 7–12 February 2020.

8. Kehl, W.; Manhardt, F.; Tombari, F.; Ilic, S.; Navab, N. SSD-6D: Making RGB-Based 3D Detection and 6D Pose Estimation Great Again. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017.

9. Tekin, B.; Sinha, S.N.; Fua, P. Real-Time Seamless Single Shot 6D Object Pose Prediction. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018.

10. Peng, S.; Liu, Y.; Huang, Q.; Zhou, X.; Bao, H. PVNet: Pixel-wise Voting Network for 6DoF Pose Estimation. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019.

11. Kao, Y.; Li, W.; Wang, Q.; Lin, Z.; Kim, W.; Hong, S. Synthetic Depth Transfer for Monocular 3D Object Pose Estimation in the Wild. In Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), New York, NY, USA, 7–12 February 2020.

12. Huang, L.; Tan, J.; Meng, J.; Liu, J.; Yuan, J. HOT-Net: Non-Autoregressive Transformer for 3D Hand-Object Pose Estimation. In Proceedings of the 28th ACM International Conference on Multimedia, Seattle, WA, USA, 12–16 October 2020.

13. Hamer, H.; Schindler, K.; Koller-Meier, E.; Van Gool, L. Tracking a Hand Manipulating an Object. In Proceedings of the 2009 IEEE International Conference on Computer Vision (ICCV), Kyoto, Japan, 29 September–2 October 2009.

14. Choi, C.; Ho Yoon, S.; Chen, C.-N.; Ramani, K. Robust Hand Pose Estimation during the Interaction with an Unknown Object. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017.
15. 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 Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018.

16. Hasson, Y.; Varol, G.; Tzionas, D.; Kalevatyk, I.; Black, M.J.; Laptev, I.; Schmid, C. Learning Joint Reconstruction of Hands and Manipulated Objects. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019.

17. Oberweger, M.; Wohlhart, P.; Lepetit, V. Generalized feedback loop for joint hand-object pose estimation. IEEE Trans. Pattern Anal. Mach. Intell. 2019, 42, 1898–1912. [CrossRef] [PubMed]

18. Tekin, B.; Bogo, F.; Pollefeys, M. H+O: Unified Egocentric Recognition of 3D Hand-Object Poses and Interactions. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019.

19. Doosti, B.; Naha, S.; Mirbagheri, M.; Crandall, D.J. HOPE-Net: A Graph-based Model for Hand-Object Pose Estimation. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 16–18 June 2020.

20. Cai, Y.; Ge, L.; Liu, J.; Cai, J.; Cham, T.-J.; Yuan, J.; Thalmann, N.M. Exploiting Spatial-temporal Relationships for 3D Pose Estimation via Graph Convolutional Networks. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea, 27 October–3 November 2019.

21. Sridhar, S.; Mueller, F.; Zollhöfer, M.; Casas, D.; Oulasvirta, A.; Theobalt, C. Real-Time Joint Tracking of a Hand Manipulating an Object from RGB-D Input. In Proceedings of the 2016 European Conference on Computer Vision (ECCV), Amsterdam, The Netherlands, 8–16 October 2016.

22. Tsoli, A.; Argyros, A.A. Joint 3D Tracking of a Deformable Object in Interaction with a Hand. In Proceedings of the 2018 European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018.

23. Baek, S.; Kim, K.I.; Kim, T.-K. Weakly-supervised Domain Adaptation via GAN and Mesh Model for Estimating 3D Hand Poses Interacting Objects. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 16–18 June 2020.

24. Romero, J.; Tzionas, D.; Black, M.J. Embodied hands: Modeling and capturing hands and bodies together. ACM Trans. Graph. 2017, 36, 1–17. [CrossRef]

25. Kipf, T.N.; Welling, M. Semi-supervised Classification with Graph Convolutional Networks. In Proceedings of the 5th International Conference on Learning Representations (ICLR), Toulon, France, 24–26 April 2017.

26. Li, R.; Wang, S.; Zhu, F.; Huang, J. Adaptive Graph Convolutional Neural Networks. In Proceedings of the 32nd AAAI Conference on Artificial Intelligence (AAAI), New Orleans, LA, USA, 2–7 February 2018.

27. Niepert, M.; Ahmed, M.; Kutzkov, K. Learning Convolutional Neural Networks for Graphs. In Proceedings of the 33rd International Conference on Machine Learning (ICML), New York, NY, USA, 19–24 June 2016.

28. Zhao, L.; Peng, X.; Tian, Y.; Kapadia, M.; Metaxas, D.N. Semantic Graph Convolutional Networks for 3D Human Pose Regression. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019.

29. Shi, L.; Zhang, Y.; Cheng, J.; Lu, H. Two-Stream Adaptive Graph Convolutional Networks for Skeleton-Based Action Recognition. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019.

30. Wei, S.-E.; Ramakrishna, V.; Kanade, T.; Sheikh, Y. Convolutional Pose Machines. In Proceedings of the 2016 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016.

31. Zhou, X.; Huang, Q.; Sun, X.; Xue, X.; Wei, Y. Towards 3D Human Pose Estimation in the Wild: A Weakly-supervised Approach. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017.

32. 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 Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018.

33. Hampali, S.; Rad, M.; Oberweger, M.; Lepetit, V. HOOnnotate: A method for 3D Annotation of Hand and Object Poses. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 16–18 June 2020.

34. Chang, A.X.; Funkhouser, T.; Guibas, L.; Hanrahan, P.; Huang, Q.; Li, Z.; Savarese, S.; Savva, M.; Song, S.; Su, H.; et al. Shapenet: An information-rich 3d model repository. arXiv 2015, arXiv:1512.03012.

35. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016.