Geometric Features Informed Multi-person Human-object Interaction Recognition in Videos

Tanqiu Qiao[1], Qianhui Men[2], Frederick W. B. Li[1], Yoshiki Kubotani[3], Shigeo Morishima[3], and Hubert P. H. Shum[1]†

[1] Durham University, United Kingdom
{tanqiu.qiao, frederick.li, hubert.shum}@durham.ac.uk
[2] University of Oxford, United Kingdom qianhui.men@eng.ox.ac.uk
[3] Waseda Research Institute for Science and Engineering, Japan
yoshikikubotani@akane.waseda.jp, shigeo@waseda.jp

Abstract. Human-Object Interaction (HOI) recognition in videos is important for analyzing human activity. Most existing work focusing on visual features usually suffer from occlusion in the real-world scenarios. Such a problem will be further complicated when multiple people and objects are involved in HOIs. Consider that geometric features such as human pose and object position provide meaningful information to understand HOIs, we argue to combine the benefits of both visual and geometric features in HOI recognition, and propose a novel Two-level Geometric feature-informed Graph Convolutional Network (2G-GCN). The geometric-level graph models the interdependency between geometric features of humans and objects, while the fusion-level graph further fuses them with visual features of humans and objects. To demonstrate the novelty and effectiveness of our method in challenging scenarios, we propose a new multi-person HOI dataset (MPHOI-72). Extensive experiments on MPHOI-72 (multi-person HOI), CAD-120 (single-human HOI) and Bimanual Actions (two-hand HOI) datasets demonstrate our superior performance compared to state-of-the-arts.

Keywords: Human-object interaction, graph convolution neural networks, feature fusion, multi-person interaction

1 Introduction

The real-world human activities are often closely associated with surrounding objects. Human-Object Interaction (HOI) recognition focuses on learning and analyzing the interaction between human and object entities for activity recognition. HOI recognition involves the segmentation and recognition of individual human sub-activities/object affordances in videos, such as drinking and placing, to gain an insight of the overall human activities [37]. Based on this, downstream applications such as security surveillance, healthcare monitoring and human-robot interactions can be developed.

† Corresponding author

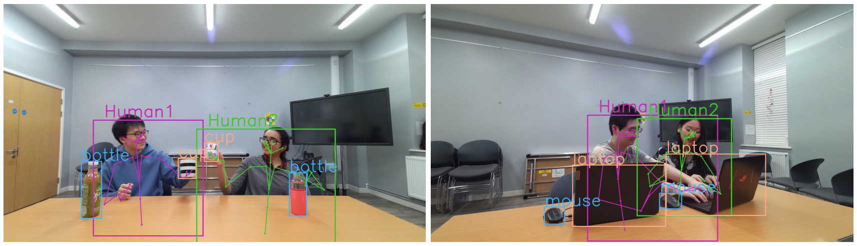

Fig. 1. Two examples (*Cheering* and *Co-working*) of our collected multi-person HOI dataset. Geometric features such as skeletons and bounding boxes are annotated.

Earlier work in HOI detection is limited to detecting interactions in one image [15,32,13]. With HOI video datasets proposed, models have been developed to learn the action representations over the spatio-temporal domain for HOI recognition [38,20]. Notably, [37] proposes a visual feature attention model to learn asynchronous and sparse HOI in videos, achieving state-of-the-art results.

A main challenge of video-based HOI recognition is that visual features usually suffer from occlusion. This is particularly problematic in real-world scenarios when multiple people and objects are involved. Recent research has shown that extracted pose features are more robust to partial occlusions than visual features [55,39]. Bottom-up pose estimators can extract body poses as long as the local image patches of joints are not occluded [4]. With advanced frameworks such as Graph Convolutional Networks (GCN), geometric pipelines generally perform better than visual ones on datasets with heavy occlusion [8]. Therefore, geometric features provide complementary information to visual ones [3,39].

In this paper, we propose to fuse geometric and visual features for HOI recognition in videos. Our research insight is that geometric features enrich fine-grained human-object interactions, as evidenced by previous research on image-based HOI detection [59,30]. We present a novel Two-level Geometric feature informed Graph Convolutional Network (2G-GCN) that extracts geometric features and fuses them with visual ones for HOI recognition in videos. We implement the network by using the geometric-level graph to model representative geometric features among humans and objects, and fusing the visual features through the fusion-level graph.

To showcase the effectiveness of our model, we further propose a multi-person dataset for Human-Object Interaction (MPHOI), which closely ensembles real-world activities that contain multiple people interacting with multiple objects. Our dataset includes common multi-person activities and natural occlusions in daily life (Fig. 1). It is annotated with the geometric features of human skeletal poses, human and object bound boxes, and ground-truth HOI activity labels, which can be used as a versatile benchmark for multiple tasks such as visual-based or skeleton-based human activity analysis or hybrid.

We outperform state-of-the-arts in multiple datasets, including our novel MPHOI-72 dataset, the single-human HOI CAD-120 [24] dataset, and the two-hand Bimanual Actions [9] dataset. We also extensively evaluate core components of 2G-GCN in ablation studies. Our main contributions are as follows:

- We propose a novel geometry-informed 2G-GCN network for HOI recognition in videos. The network consists of a two-level graph structure that models geometric features between human and object, together with the corresponding visual features.
- We present the novel problem of MPHOI in videos with a new MPHOI-72 dataset, showcasing new challenges that cannot be directly resolved by existing methods. The source code and dataset are made public[1].
- We outperform state-of-the-art HOI recognition networks in our MPHOI-72 dataset, the CAD-120 [24] dataset and the Bimanual Actions [9] dataset.

2 Related Work

2.1 HOI Detection in Images

HOI detection aims at understanding interactions between humans and objects and identifying their interdependencies within a single image. Gupta and Malik [18] first address the HOI detection task, which entails recognising human activities and the object instances they interact with in an image. Assigning distinct semantic responsibilities to items in a HOI process allows a detailed understanding of the present state of affairs. Gkioxari *et al.* [14] apply an action-specific density map over target object locations depending on the appearance of an identified person to the system in [18]. Multiple large-scale datasets have been presented in recent years for exploring HOI detection in images, such as V-COCO [18], HICO-DET [5] and HCVRD [61]. Specifically, Mallya and Lazebnik [32] present a simple network that fuses characteristics from a human bounding box and the entire image to detect HOIs. Gao *et al.* [13] exploit an instance-centric attention module to improve the information from regions of interest and assist HOI classification. These early methods focus on visual relationships between entities in images without any potential structural relationships in HOIs.

Graph Convolutional Networks (GCN) [22] can be used to assimilate valuable expressions of graph-structured data. Kato *et al.* [21] employ it to assemble new HOIs by using information from WordNet [35]. Xu *et al.* [56] also exploit a GCN to model the semantic dependencies between action and object categories. Wang *et al.* [51] hypothesise that it is convenient to represent the entities as nodes and the relations as the edges connecting them in HOI. They design a contextual heterogeneous graph network to deeply explore the relations between people and objects. VSGNet [48] refines the visual features from human-object pairs with the spatial configuration, and exploits the structural connections between pairs through graph convolution. These approaches achieve remarkable performance in image data and can provide a basis for HOI recognition in videos.

[1] `https://github.com/tanqiu98/2G-GCN`

2.2 HOI Recognition in Videos

HOI recognition in videos requires high-level spatial and temporal reasoning between humans and objects. Some earlier attempts apply spatio-temporal context to achieve rich-context for HOI recognition [29,24,17]. Recent works combine graphical models with deep neural networks (DNNs). Jain *et al.* [20] propose a model for integrating the strength of spatio-temporal graphs with Recurrent Neural Networks (RNNs) in sequence learning. Qi *et al.* [38] expand prior graphical models in DNNs for videos with learnable graph structures and pass messages through GPNN. Dabral *et al.* [6] analyze the effectiveness of GCNs against Convolutional Networks and Capsule Networks for spatial relation learning. Wang *et al.* [53] propose the STIGPN exploiting the parsed graphs to learn spatio-temporal connection development and discover objects existing in a scene. Although previous methods attain impressive improvements in specific tasks, they are all based on visual features, which are unreliable in real-life HOI activities that contain occlusions between human and object entities.

2.3 HOI Recognition Datasets

Multiple datasets are available to research HOI in videos for different tasks. CAD-120 [24], Bimanual Actions [9], Bimanual Manipulation [25], *etc.* are useful for single-person HOI recognition. The latter two also present bimanual HOI recognition tasks as they record human activities using both hands for object interaction. Something-Else [34], VLOG [12], EPIC Kitchens [7] are available for single-hand HOI recognition tasks, where the EPIC Kitchens dataset can be also used for bimanual HOI recognition since it captures both hands during cooking. The UCLA HHOI Dataset [46,45] focuses on human-human-object interaction, involving at most two humans and one object. As true multi-person HOI should involve multiple humans and objects, we propose a novel MPHOI dataset that collects daily activities with multiple people interacting multiple objects.

2.4 Geometric Features informed HOI Analysis

Recent research begins to employ human pose to the HOI tasks in images, which takes the advantage of capturing structured connections in human skeletons. To focus on the important aspects of interaction, Fang *et al.* [10] suggest a pairwise body-part attention model. Based on semantic attention, Wan *et al.* [50] provide a zoom-in module for extracting local characteristics of human body joints. Zheng *et al.* [59] introduce a skeleton-based interactive graph network (SIGN) to capture fine-grained HOI between keypoints in human skeletons and objects.

However, introducing geometric features such as the keypoints of human pose and objects to HOI learning in videos is challenging and underexplored for a few reasons. On the one hand, in a video, interaction definitions might be ambiguous, such as lift a cup vs. place a cup, approaching vs. retreating vs. reaching. These actions might be detected as the same image label due to their visual similarity.

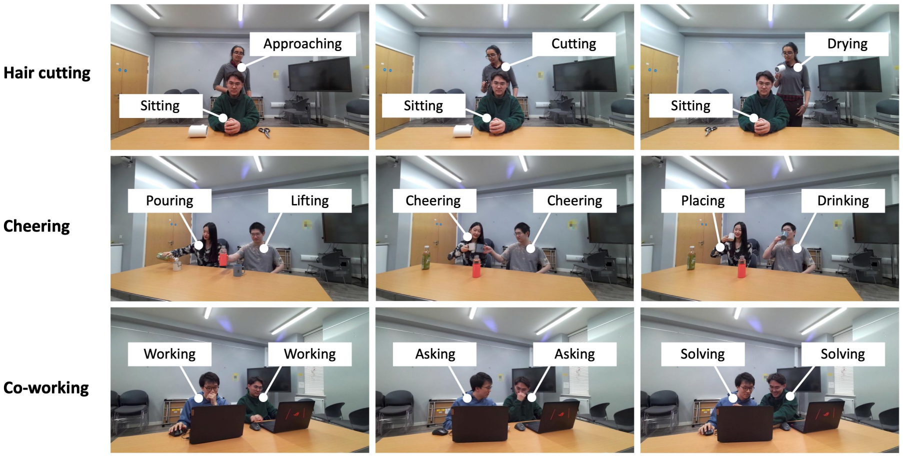

Fig. 2. Sample video frames of three different MPHOI activities in MPHOI-72.

Videos allow the use of temporal visual cues that are not presented in images [37]. On the other hand, the model needs to consider human dynamics in the video and the shifting orientations of items in the scene in relation to humans [38]. This makes it difficult to directly extend image-based models to video that exploit the region of interest (ROI) features of human-object union [6]. We propose a novel two-level graph to refine the interactive representations; the first graph models the interdependency within the geometric key points of human and objects, and the second graph models the interdependency between the visual features and the learned geometric representations.

3 The Multi-Person HOI Dataset (MPHOI-72)

We propose a HOI dataset with multi-person activities (MPHOI-72), which is challenging due to many body occlusions among the humans and objects. We have 3 males and 2 females, aged 23-27, who are randomly combined into 8 groups with 2 people per group and perform 3 different HOI activities interacting with 2-4 objects. We also prepared 6 objects: cup, bottle, scissors, hair dryer, mouse and laptop. 3 activities = {*Cheering*, *Hair cutting*, *Co-working*} and 13 sub-activities = {*Sit*, *Approach*, *Retreat*, *Place*, *Lift*, *Pour*, *Drink*, *Cheers*, *Cut*, *Dry*, *Work*, *Ask*, *Solve*} are defined. The dataset consists of 72 videos captured from 3 different angles at 30 fps, with totally 26,383 frames and an average length of 12 seconds.

Fig. 2 shows some sample video frames of the three activities in our MPHOI-72 dataset, and the sub-activity label of each subject is annotated frame-wise. The top row presents *Hair cutting* from the front view, where one subject is sitting and another subject interacts with a pair of scissors and a hair dryer. Most part of the body of the subject standing at the back is invisible. The second row presents a popular human activity, *Cheering*, in which two subjects pour water from their own bottles, lift cups to cheer, and drink. The high-level occlusion exists between humans, cups and bottles during the entire activity. The bottom row presents *Co-working*, which simulates the situation of two co-workers asking and solving questions. Besides, we also consider distinct human sizes, skin colors and a balance of gender. These samples illustrate the diversity of our dataset.

We use Azure Kinect SDK to collect RGB-D videos with 3840 × 2160 resolution, and employ their Body Tracking SDK [1] to capture the full dynamics of two subject skeletons. Object bounding boxes are manually annotated framewise. For each video, we provide such geometric features: 2D human skeletons and bounding boxes of the subjects and objects involved in the activity (Fig. 1).

4 Two-level Geometric Features Informed Graph Convolutional Network (2G-GCN)

To learn the correlations during human-object interaction, we propose a two-level graph structure to model the interdependency of the geometric features, known as 2G-GCN. The model consists of two key components: a geometry-level graph for modeling geometry and object features to facilitate graph convolution learning, and a fusion-level graph for fusing geometric and visual features (Fig. 3).

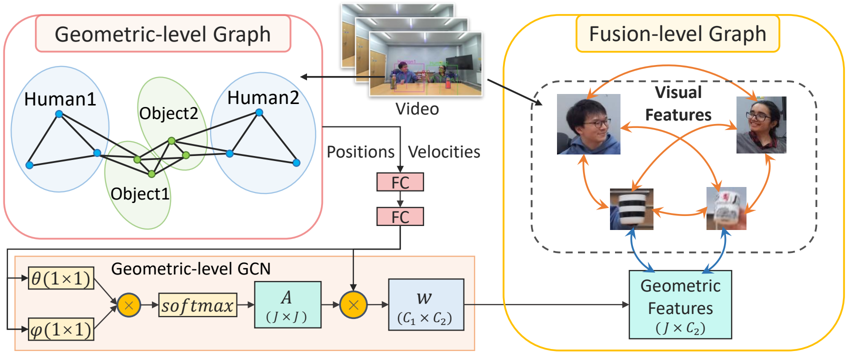

Fig. 3. Our 2G-GCN framework comprises a geometric-level and a fusion-level graph.

4.1 Geometric Features

The geometric features of humans can be represented in various ways. Human skeletons contain an explicit graph structure with joints as nodes and bones as edges. The joint position and velocity offer fine-grained dynamics in the human motion [58], while the joint angle also provides spatial cues in 3D skeleton data [43]. Alternatively, body shapes and how they deform during movement can be represented by surface models [31] or implicit models [41]. We employ human skeletons with joint position and velocity, because they are essential cues to human motion. Also, unlike body shapes, they are invariant to human appearance.

We represent human poses in an effective representation to inform HOI recognition. For human skeleton, we select specific body keypoints and denote them as a set $\mathcal{S} = \{M_t^{h,k}\}_{t=1,h=1,k=1}^{T,H,K}$, where $M_t^{h,k}$ denotes the body joint of type k in human h at time t, T denotes the total number of frames in the video, H and K denote the total number of humans and keypoints of a human body in a frame, respectively. For a given human body keypoint $M_t^{h,k}$, we define its position as $\mathbf{p}_{t,h,k} = (x_{t,h,k}, y_{t,h,k})^T \in \mathbb{R}^2$ in 2D, and the velocity as $\mathbf{v}_{t,h,k} = \mathbf{p}_{t+1,h,k} - \mathbf{p}_{t,h,k}$, which is the forward difference of neighbour frames. In the channel of each human skeleton keypoint, we concatenate its position $\mathbf{p}_{t,h,k}$, and velocity $\mathbf{v}_{t,h,k}$ in the channel domain, forming the human geometric context $\mathbf{h}_{t,h,k} = [\mathbf{p}_{t,h,k}, \mathbf{v}_{t,h,k}] \in \mathbb{R}^4$.

As objects play a crucial role in the HOI videos, we also consider their geometric features. The two diagonal points of the object bounding box are utilised to represent the object position. We define all object keypoints as $\mathcal{O} = \{B_t^{f,u}\}_{t=1,f=1,u=1}^{T,F,2}$, where $B_t^{f,u}$ denotes the object keypoint of type u in object f at time t. F denotes the maximum number of objects in a video and $u = \{1, 2\}$ is the index of the top-left and the bottom-right points of the object bounding box, respectively. The object geometric context $\mathbf{o}_{t,f,u} = [\mathbf{p}_{t,f,u}, \mathbf{v}_{t,f,u}] \in \mathbb{R}^4$ can be obtained by the same process as the human skeleton.

4.2 The Geometric-level Graph

We design a novel geometric-level graph that involves both human skeleton and object keypoints to explore their correlations in an activity (Fig. 3 left). We use $\mathbf{g}_t$ to denote a graph node with geometric features of a keypoint from either a human $h_{t,h,k}$ or a object $o_{t,f,u}$ at frame t. Therefore, all keypoints of the frame t are denoted by $G_t = (\mathbf{g}_{t,1}; \cdots; \mathbf{g}_{t,J})$, where $J = H \times K + F \times 2$ joints, each with 4 channel dimensions including its 2D position and velocity. This enables us to enhance the ability of GCN to capture correlations between human and object keypoints in HOI activities by learning their dynamic spatial cues. We embed $\mathbf{g}_t$ using two fully connected (FC) layers following [58] as:

$$\widetilde{\mathbf{g}_t} = \sigma(W_2(\sigma(W_1\mathbf{g}_t + \mathbf{b}_1)) + \mathbf{b}_2) \in \mathbb{R}^{C_1}, \tag{1}$$

where C_1 is the dimension of the joint representation, $W_1 \in \mathbb{R}^{C_1 \times 4}$ and $W_2 \in \mathbb{R}^{C_1 \times C_1}$ are weight matrices, $\mathbf{b}_1$ and $\mathbf{b}_2$ are the bias vectors, and σ is the ReLU activation function.

We propose an adaptive adjacency matrix exploiting the similarity of the geometric features in the GCN. We employ the dot-product similarity in $\widetilde{\mathbf{g}_t}$, as it allows us to determine if and how strong a connection exists between two keypoints in the same frame t [54,43,58]. This is a better choice for our problem comparing to other strategies, *e.g.* the traditional adjacency matrix only represents the physical structure of the human body [57] or a fully-learned adjacency matrix without supervision of graph representations [43]. We represent the adjacency matrix A_t with ${j_1}^{th}$ and ${j_2}^{th}$ keypoints as:

$$A_t(j_1, j_2) = \theta(\widetilde{\mathbf{g}}_{t,j_1})^T \phi(\widetilde{\mathbf{g}}_{t,j_2}), \tag{2}$$

where $\theta, \phi \in \mathbb{R}^{C_2}$ denote two transformation functions, each implemented by a 1×1 convolutional layer. Then, SoftMax activation is conducted on each row of A_t to ensure the integration of all edge weights of a node equal to 1. We subsequently obtain the output of the geometry-level graph from the GCN as:

$$Y_t = A_t \widetilde{\mathbf{G}}_t W_g, \tag{3}$$

where $\widetilde{\mathbf{G}}_t = (\widetilde{\mathbf{g}}_{t,1}; \cdots; \widetilde{\mathbf{g}}_{t,J}) \in \mathbb{R}^{J \times C_1}$ and $W_g \in \mathbb{R}^{C_1 \times C_2}$ is the transformation matrix. The size of output is $T \times J \times C_2$.

4.3 The Fusion-Level Graph

We propose a fusion-level graph to connect the geometric features learned from GCN with visual features. Previous works on CNN-based HOI recognition in videos overemphasise visual features and neglect geometric features of humans and objects [33,27]. State-of-the-arts like ASSIGN [37] also exclude geometric features. In contrast, we first extract visual features for each human or object entity by ROI pooling, and then introduce the geometric output Y_t from the GCN as the auxiliary feature to complement the visual representation. The feature vectors for all entities are then embedded by a two-layer MLP with ReLU activation function to the same hidden size.

A key design of the fusion-level graph is an attention mechanism to estimate the relevance of the interacted neighbouring entity. As illustrated in the fusion-level graph of Fig. 3, each person and object denote an entity through the time, while Y_t forms an additional entity joining the graph. All connections between the visual features of all humans and objects in the video are captured, represented by orange arrows. The blue arrows denote the connection between geometric and object visual features. Empirically, connecting the geometry-object pairs consistently performs better than applying a fully-connected graph with geometry-human connections. A possible reason is that humans are generally bigger in size and therefore have a larger chance of occlusion. Correlating such relatively noisy human visual and geometry features is a harder problem than the objects' equivalent. The fusion strategy is evaluated in the ablation studies.

The attention mechanism employed in the fusion-level graph calculates a weighted average of the contributions from neighbouring nodes, implemented by a variant of scaled dot-product attention [49] with identical keys and values:

$$\operatorname{Att}\left(q, \{z_i\}_{i=1 \ldots n}\right) = \sum_{i=1}^{n} \operatorname{softmax}\left(\frac{q^T z_i}{\sqrt{d}}\right) z_i, \tag{4}$$

where q is a query vector, $\{z_i\}$ is a set of keys/values vectors of size n, and d is the feature dimension.

Once fusion-level graph is constructed, we employ ASSIGN [37] as the backbone for HOI recognition. ASSIGN is a recurrent graph network that automatically detects the structure of HOI associated with asynchronous and sparse entities in videos. Our fusion-level graph is compatible with the HOI graph structure in ASSIGN, allowing us to employ the network to predict sub-activities for humans and object-affordances for objects depending on the dataset.

5 Experiments

5.1 Datasets

We have performed experiments on our MPHOI-72 dataset, the CAD-120 [24] dataset and the Bimanual Actions [9] dataset, showcasing the superior results of 2G-GCN on multi-person, single-human and two-hand HOI recognition.

CAD-120 is widely used for HOI recognition. It consists of 120 RGB-D videos of 10 different activities performed individually by 4 participants, with each activity replicated 3 times. A participant interacts with 1-5 objects in each video. There are 10 human sub-activities (*e.g.*, *eating*, *drinking*), and 12 object affordances (*e.g.*, *stationary*, *drinkable*) in total, which are annotated per frame.

Bimanual Actions is the first HOI activity dataset where subjects use two hands to interact with objects (*e.g.*, the left hand holding a piece of wood, while the right hand sawing it). It contains 540 RGB-D videos of 6 subjects performing 9 different activities, with each repeated for 10 times. There are totally 14 action labels for each hand and each entity in a video is annotated frame-wise.

5.2 Implementation Details

Network Settings We implement 2048-dimensional ROI pooling features extracted from the 2D bounding boxes of humans and objects in the video detected by a Faster R-CNN [40] module, which is pre-trained [2] on the Visual Genome dataset [26] for entity visual features. We set the number of neurons to 64, 128 for both FC layers for the embedding and the transformation functions of Eq. 2 in the geometric-level graph, respectively (*i.e.*, $C_1 = 64$, $C_2 = 128$).

Experimental Settings 2G-GCN is evaluated on two tasks: 1) joined segmentation, and 2) label recognition given known segmentation. The first task needs the model to segment and identify the timeline for each entity in a video. The second task is a variant of the previous one, in which the ground-truth segmentation is known and the model requires to name the existing segments. For the Bimanual Actions and CAD-120 datasets, we use leave-one-subject cross-validation to evaluate the generalization effort of 2G-GCN in unknown subjects. On MPHOI-72, we define a cross-validation scheme that chooses two subjects not present in the training set as the test set.

For evaluation, we report the F$_1$@k metric [28] with the commonly used thresholds $k = 10\%$, 25% and 50%. The F$_1$@k metric believes each predicted action segment is correct if its Intersection over Union (IoU) ratio with respect to the corresponding ground truth is at least k. Since it is more sensitive to short action classes and over-segmentation errors, F$_1$ is more adaptable than the frame-based metrics for joined segmentation and labelling issues, and was frequently adopted in prior segmentation researches [28,11,37].

With four Nvidia Titan RTX GPUs, training MPHOI-72, CAD-120 and Bimanual Actions takes 2 hours, 8 hours and 5 days, respectively. Testing the whole test set takes 2 minutes, 3 minutes and 20 minutes, respectively.

5.3 Quantitative Comparison

Multi-person HOIs In our challenging MPHOI-72 dataset, 2G-GCN beats ASSIGN [37] by a considerable gap (Table 1). 2G-GCN significantly outperforms ASSIGN and has smaller standard deviation values in every F$_1$ configurations, reaching 68.6% in F$_1$@10 score, which is approximately 9.5% higher than ASSIGN. The performance of visual-based methods such as ASSIGN is generally ineffective, since remarkable occlusions in MPHOI typically invalids visual features to HOI recognition task. The significant gaps between the results of 2G-GCN and ASSIGN demonstrate that the application of geometric features and its fusion with visual features can motivate our model to learn stable and essential features even when significant occlusion appears in HOIs.

Table 1. Joined segmentation and label recognition on MPHOI-72.

Model	Sub-activity		
	F$_1$@10	F$_1$@25	F$_1$@50
ASSIGN [37]	59.1 ± 12.1	51.0 ± 16.7	33.2 ± 14.0
2G-GCN	**68.6 ± 10.4**	**60.8 ± 10.3**	**45.2 ± 6.5**

Table 2. Joined segmentation and label recognition on CAD-120.

Model	Sub-activity			Object Affordance		
	F$_1$@10	F$_1$@25	F$_1$@50	F$_1$@10	F$_1$@25	F$_1$@50
rCRF [42]	65.6 ± 3.2	61.5 ± 4.1	47.1 ± 4.3	72.1 ± 2.5	69.1 ± 3.3	57.0 ± 3.5
Independent BiRNN	70.2 ± 5.5	64.1 ± 5.3	48.9 ± 6.8	84.6 ± 2.1	81.5 ± 2.7	71.4 ± 4.9
ATCRF [23]	72.0 ± 2.8	68.9 ± 3.6	53.5 ± 4.3	79.9 ± 3.1	77.0 ± 4.1	63.3 ± 4.9
Relational BiRNN	79.2 ± 2.5	75.2 ± 3.5	62.5 ± 5.5	82.3 ± 2.3	78.5 ± 2.7	68.9 ± 4.9
ASSIGN [37]	88.0 ± 1.8	84.8 ± 3.0	73.8 ± 5.8	92.0 ± 1.1	90.2 ± 1.8	82.4 ± 3.5
2G-GCN	**89.5 ± 1.6**	**87.1 ± 1.8**	**76.2 ± 2.8**	**92.4** ± 1.7	**90.4** ± 2.3	**82.7 ± 2.9**

Table 3. Joined segmentation and label recognition on Bimanual Actions.

Model	Sub-activity		
	F_1@10	F_1@25	F_1@50
Dreher *et al.* [9]	40.6 ± 7.2	34.8 ± 7.1	22.2 ± 5.7
Independent BiRNN	74.8 ± 7.0	72.0 ± 7.0	61.8 ± 7.3
Relational BiRNN	77.7 ± 3.9	75.0 ± 4.2	64.8 ± 5.3
ASSIGN [37]	84.0 ± 2.0	81.2 ± 2.0	68.5 ± 3.3
2G-GCN	**85.0** ± 2.2	**82.0** ± 2.6	**69.2** ± **3.1**

Single-person HOIs The generic formulation of 2G-GCN results in excellent performance in single-person HOI recognition. Table 2 presents the results of 2G-GCN with state-of-the-arts and two BiRNN-based baselines on CAD-120. Bidirectional GRU is used as a baseline in both cases: The Independent BiRNN models each entity individually (*i.e.*, there are no spatial messages), but the Relational BiRNN incorporates extensive spatial relations between entities. Three previous works, ATCRF [23], rCRF [42] and ASSIGN [37], are fully capable of performing this task, where ASSIGN is relatively new and can improve the scores to higher levels. For both human sub-activity and object affordance labelling, 2G-GCN beats ASSIGN in every configuration of the $F_1@k$ metric. Especially for the sub-activity labelling, 2G-GCN improves 1.5% over ASSIGN in F_1@10, and more than 2% in $F_1@\{25, 50\}$ with lower standard deviation values. These findings demonstrate the benefits of using geometric features from human skeletons and object bounding boxes, rather than only using visual features like ASSIGN.

Two-hand HOIs For two-hand HOI recognition on the Bimanual Actions dataset, 2G-GCN outperforms ASSIGN [37] by about 1%. We compare the performance on the joined segmentation and labelling task with Dreher *et al.* [9], ASSIGN [37] and BiRNN baselines (Table 3). Dreher et al. [9] have the worst results due to their fairly basic graph network, which ignores hand interactions and does not account for long-term temporal context. By taking into account a larger temporal context, the BiRNN baselines outperform Dreher et al. [9]. Our 2G-GCN has made a small improvement over ASSIGN [37]. This is partly because the hand skeletons provided by the Bimanual Actions dataset are extracted by OpenPose [4], which is relatively weak on hand pose estimation.

5.4 Qualitative Comparison

We compare the visualization of 2G-GCN and relevant methods on our challenging MPHOI-72 dataset. Fig. 4 shows an example of segmentation and labelling results with 2G-GCN and ASSIGN [37] approaches compared with the ground-truth for a *Cheering* activity. We highlight some major segmentation errors with red dashed boxes. Although both models have some errors, 2G-GCN is generally more robust to varying segmentation period and activity progression than

ASSIGN. 2G-GCN is not particularly sensitive to the timeline of *place* and *approach*, while ASSIGN crashes for most of the activities.

Fig. 5 displays an example of a *taking food* activity on the CAD-120 dataset. We highlight over-segmentation with the red dashed box and chaotic segmentation with the blue dashed box. From the figure, our 2G-GCN is able to segment and recognise both human sub-activities and object affordances more accurately than the other two models. ASSIGN [37] and Relational BiRNN fail to predict when the human opens or closes the microwave (*e.g.* the *open* and *close* sub-activities for the human, and the *openable* and *closable* affordances for the microwave).

Fig. 6 depicts the qualitative visualization of a *cooking* activity on the Bimanual Actions dataset. Here, 2G-GCN performs outstandingly with precise segmentation and labelling results for the left hand, while ASSIGN [37] and Relational BiRNN have a chaotic performance when segmenting the long *stir* action. In contrast, the right hand has more complex actions, which confuses the models a lot. 2G-GCN generally performs better than ASSIGN, although both of them have some additional and missing segmentations. Relational BiRNN has the worst performance with chaotic segmentation errors in the *hold* action.

5.5 Ablation Studies

The two proposed graphs in our method contain important structural information. We ablate various essential modules and evaluate them on the CAD-120 dataset to demonstrate the role of different 2G-GCN components as shown in Table 4, where GG and FG denote the geometric-level graph and fusion-level graph, respectively.

We firstly investigate the importance for geometric features of the human and objects. The experiments in row (1) drops the human skeleton features in the geometric-level graph, while row (2) drops the object keypoint features. Row (3) explores the effect of the embedding function on geometric features. The last component we ablated is the similarity matrix used in the GCN, the result comparison between row (4) and (8) demonstrates its significance in the model.

We further ablate different components in the fusion-level graph as shown in Fig. 7. We disable the attention connection between the pair of human-object and

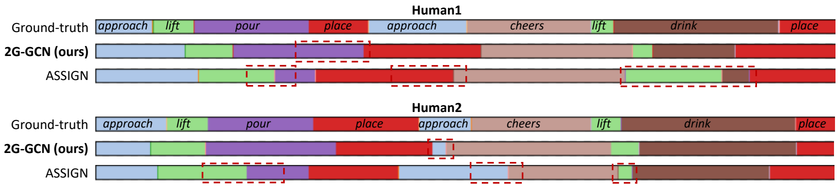

Fig. 4. Visualizing the segmentation and labels on MPHOI-72 for *Cheering*. Red dashed boxes highlights major segmentation errors.

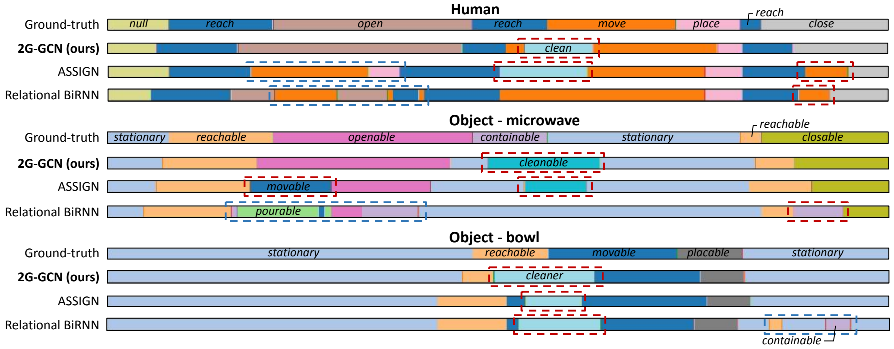

Fig. 5. Visualizing the segmentation and labels on CAD-120 for *taking food*. Red dashed boxes highlight over-segmentation. Blue ones highlight chaotic segmentation.

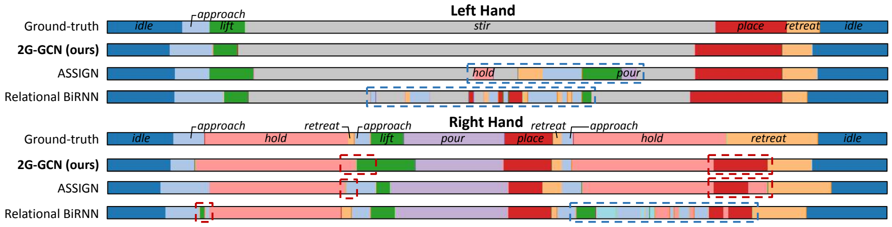

Fig. 6. Visualizing the segmentation and labels on Bimanual Actions for *cooking*. Red dashed boxes highlight extra or missing segmentation. Blue ones highlights chaotic segmentation.

object-object in row (5) and (6), respectively, and also supplement the human-geometry connection in row (7). The inferior results reported in row (5) and (6) verify the significance of incorporating all these pair connections in our full 2G-GCN model.

6 Conclusions

We propose a two-level graph GCN for tackling HOIs in videos, which consists of a geometric-level graph using human skeletons and object bounding boxes, and a fusion-level graph fusing the geometric features with traditional visual features. We also propose a novel MPHOI-72 dataset to enable and motivate research in multi-person HOI recognition. Our 2G-GCN outperforms state-of-the-art HOI recognition networks in single-person, two-hand and multi-person HOI domains.

Our method is not limited to two humans; the geometric-level graph can represent multiple humans and objects. To handle an arbitrary number of entities, a graph can be constructed by only considering the k-nearest humans and objects, allowing better generalisation [36]. If there are a large number of entities, to avoid

Table 4. Ablation study on CAD-120. GG and FG denote the geometric-level graph and the fusion-level graph, respectively.

Model	Sub-activity F$_1$@10	Sub-activity F$_1$@25	Object Affordance F$_1$@10	Object Affordance F$_1$@25
(1) GG (w/o skeletons) & FG	87.7	84.9	91.0	88.3
(2) GG (w/o objects) & FG	88.3	85.6	90.4	88.5
(3) GG (w/o embedding) & FG	89.4	86.4	91.5	90.0
(4) GG (w/o similarity) & FG	88.7	85.0	90.6	89.0
(5) GG & FG (w/o human-object)	73.4	68.8	90.3	88.4
(6) GG & FG (w/o object-object)	88.3	84.5	90.9	88.5
(7) GG & FG (w human-geometry)	89.0	86.6	91.4	89.3
(8) 2G-GCN	**89.5**	**87.1**	**92.4**	**90.4**

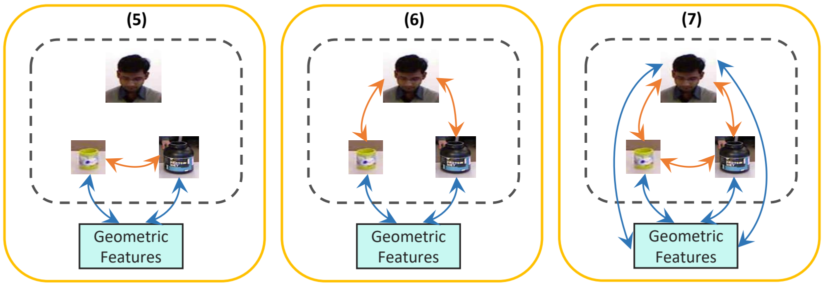

Fig. 7. Ablation study of the fusion-level graph. Human-object, object-object and geometry-human relations are ablated (rows (5), (6), (7) in Table 4 respectively).

handling a large fully-connected graph, we can apply an attention mechanism to learn what nodes are related [44], thereby better recognising HOIs.

We found that the accuracy of skeleton joint detection can affect the quality of geometric features. In future work, we may employ some algorithms for noise handling. Skeleton reconstruction methods such as the lazy learning approach [47] or motion denoising methods such as the deep learning manifold [52] would enhance the accuracy of skeleton information based on prior learning from a dataset of natural motion. One of our future directions is to employ such techniques to improve our geometric features.

Another future direction is to enrich the geometric representation of objects. While the bounding box features are powerful, it cannot represent the geometric details [60]. On the one hand, the rotation-equivariant detector [19] enriches the object representation with rotated bounding boxes, resulting in improved object detection performance. On the other hand, the recently proposed convex-hull features [16] allow representing objects of irregular shapes and layouts. They could enhance our geometric feature based HOI recognition framework significantly.

References

1. Quickstart: Set up azure kinect body tracking (2022), `https://docs.microsoft.com/en-us/azure/kinect-dk/body-sdk-setup`
2. Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., Zhang, L.: Bottom-up and top-down attention for image captioning and visual question answering. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 6077–6086 (2018)
3. Bodla, N., Shrivastava, G., Chellappa, R., Shrivastava, A.: Hierarchical video prediction using relational layouts for human-object interactions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12146–12155 (2021)
4. Cao, Z., Hidalgo, G., Simon, T., Wei, S.E., Sheikh, Y.: Openpose: Realtime multi-person 2d pose estimation using part affinity fields. arXiv e-prints pp. arXiv–1812 (2018)
5. Chao, Y.W., Liu, Y., Liu, X., Zeng, H., Deng, J.: Learning to detect human-object interactions. In: 2018 ieee winter conference on applications of computer vision (wacv). pp. 381–389. IEEE (2018)
6. Dabral, R., Sarkar, S., Reddy, S.P., Ramakrishnan, G.: Exploration of spatial and temporal modeling alternatives for hoi. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 2281–2290 (2021)
7. Damen, D., Doughty, H., Farinella, G.M., , Furnari, A., Ma, J., Kazakos, E., Moltisanti, D., Munro, J., Perrett, T., Price, W., Wray, M.: Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100. International Journal of Computer Vision (IJCV) (2021), `https://doi.org/10.1007/s11263-021-01531-2`
8. Das, S., Sharma, S., Dai, R., Bremond, F., Thonnat, M.: Vpn: Learning video-pose embedding for activities of daily living. In: European Conference on Computer Vision. pp. 72–90. Springer (2020)
9. Dreher, C.R., Wächter, M., Asfour, T.: Learning object-action relations from bimanual human demonstration using graph networks. IEEE Robotics and Automation Letters **5**(1), 187–194 (2020)
10. Fang, H.S., Cao, J., Tai, Y.W., Lu, C.: Pairwise body-part attention for recognizing human-object interactions. In: Proceedings of the European conference on computer vision (ECCV). pp. 51–67 (2018)
11. Farha, Y.A., Gall, J.: Ms-tcn: Multi-stage temporal convolutional network for action segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3575–3584 (2019)
12. Fouhey, D.F., Kuo, W.c., Efros, A.A., Malik, J.: From lifestyle vlogs to everyday interactions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4991–5000 (2018)
13. Gao, C., Zou, Y., Huang, J.B.: ican: Instance-centric attention network for human-object interaction detection. arXiv preprint arXiv:1808.10437 (2018)
14. Gkioxari, G., Girshick, R., Dollár, P., He, K.: Detecting and recognizing human-object interactions. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 8359–8367 (2018)
15. Gkioxari, G., Girshick, R., Malik, J.: Actions and attributes from wholes and parts. In: Proceedings of the IEEE international conference on computer vision. pp. 2470–2478 (2015)

16. Guo, Z., Liu, C., Zhang, X., Jiao, J., Ji, X., Ye, Q.: Beyond bounding-box: Convex-hull feature adaptation for oriented and densely packed object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 8792–8801 (June 2021)
17. Gupta, A., Kembhavi, A., Davis, L.S.: Observing human-object interactions: Using spatial and functional compatibility for recognition. IEEE transactions on pattern analysis and machine intelligence **31**(10), 1775–1789 (2009)
18. Gupta, S., Malik, J.: Visual semantic role labeling. arXiv preprint arXiv:1505.04474 (2015)
19. Han, J., Ding, J., Xue, N., Xia, G.S.: Redet: A rotation-equivariant detector for aerial object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2786–2795 (June 2021)
20. Jain, A., Zamir, A.R., Savarese, S., Saxena, A.: Structural-rnn: Deep learning on spatio-temporal graphs. In: Proceedings of the ieee conference on computer vision and pattern recognition. pp. 5308–5317 (2016)
21. Kato, K., Li, Y., Gupta, A.: Compositional learning for human object interaction. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 234–251 (2018)
22. Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907 (2016)
23. Koppula, H.S., Saxena, A.: Anticipating human activities using object affordances for reactive robotic response. IEEE transactions on pattern analysis and machine intelligence **38**(1), 14–29 (2016)
24. Koppula, H.S., Gupta, R., Saxena, A.: Learning human activities and object affordances from rgb-d videos. The International Journal of Robotics Research **32**(8), 951–970 (2013)
25. Krebs, F., Meixner, A., Patzer, I., Asfour, T.: The kit bimanual manipulation dataset. In: IEEE/RAS International Conference on Humanoid Robots (Humanoids). pp. 0–0 (2021)
26. Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.J., Shamma, D.A., et al.: Visual genome: Connecting language and vision using crowdsourced dense image annotations. International journal of computer vision **123**(1), 32–73 (2017)
27. Le, H., Sahoo, D., Chen, N.F., Hoi, S.C.: Bist: Bi-directional spatio-temporal reasoning for video-grounded dialogues. arXiv preprint arXiv:2010.10095 (2020)
28. Lea, C., Flynn, M.D., Vidal, R., Reiter, A., Hager, G.D.: Temporal convolutional networks for action segmentation and detection. In: proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 156–165 (2017)
29. Li, Y., Nevatia, R.: Key object driven multi-category object recognition, localization and tracking using spatio-temporal context. In: European Conference on Computer Vision. pp. 409–422. Springer (2008)
30. Liang, Z., Liu, J., Guan, Y., Rojas, J.: Pose-based modular network for human-object interaction detection. arXiv preprint arXiv:2008.02042 (2020)
31. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: Smpl: A skinned multi-person linear model. ACM transactions on graphics (TOG) **34**(6), 1–16 (2015)
32. Mallya, A., Lazebnik, S.: Learning models for actions and person-object interactions with transfer to question answering. In: European Conference on Computer Vision. pp. 414–428. Springer (2016)

33. Maraghi, V.O., Faez, K.: Zero-shot learning on human-object interaction recognition in video. In: 2019 5th Iranian conference on signal processing and intelligent systems (ICSPIS). pp. 1–7. IEEE (2019)
34. Materzynska, J., Xiao, T., Herzig, R., Xu, H., Wang, X., Darrell, T.: Something-else: Compositional action recognition with spatial-temporal interaction networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1049–1059 (2020)
35. Miller, G.A.: Wordnet: a lexical database for english. Communications of the ACM **38**(11), 39–41 (1995)
36. Mohamed, A., Qian, K., Elhoseiny, M., Claudel, C.: Social-stgcnn: A social spatio-temporal graph convolutional neural network for human trajectory prediction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14424–14432 (2020)
37. Morais, R., Le, V., Venkatesh, S., Tran, T.: Learning asynchronous and sparse human-object interaction in videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16041–16050 (2021)
38. Qi, S., Wang, W., Jia, B., Shen, J., Zhu, S.C.: Learning human-object interactions by graph parsing neural networks. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 401–417 (2018)
39. Qiu, L., Zhang, X., Li, Y., Li, G., Wu, X., Xiong, Z., Han, X., Cui, S.: Peeking into occluded joints: A novel framework for crowd pose estimation. In: European Conference on Computer Vision. pp. 488–504. Springer (2020)
40. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: towards real-time object detection with region proposal networks. IEEE transactions on pattern analysis and machine intelligence **39**(6), 1137–1149 (2016)
41. Saito, S., Simon, T., Saragih, J., Joo, H.: Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2020)
42. Sener, O., Saxena, A.: rcrf: Recursive belief estimation over crfs in rgb-d activity videos. In: Robotics: Science and systems. Citeseer (2015)
43. Shi, L., Zhang, Y., Cheng, J., Lu, H.: Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12026–12035 (2019)
44. Shi, L., Wang, L., Long, C., Zhou, S., Zhou, M., Niu, Z., Hua, G.: Sgcn: Sparse graph convolution network for pedestrian trajectory prediction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8994–9003 (2021)
45. Shu, T., Gao, X., Ryoo, M.S., Zhu, S.C.: Learning social affordance grammar from videos: Transferring human interactions to human-robot interactions. In: 2017 IEEE international conference on robotics and automation (ICRA). pp. 1669–1676. IEEE (2017)
46. Shu, T., Ryoo, M.S., Zhu, S.C.: Learning social affordance for human-robot interaction. arXiv preprint arXiv:1604.03692 (2016)
47. Shum, H.P., Ho, E.S., Jiang, Y., Takagi, S.: Real-time posture reconstruction for microsoft kinect. IEEE transactions on cybernetics **43**(5), 1357–1369 (2013)
48. Ulutan, O., Iftekhar, A.S.M., Manjunath, B.S.: Vsgnet: Spatial attention network for detecting human object interactions using graph convolutions. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 13617–13626 (2020)

49. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in neural information processing systems. pp. 5998–6008 (2017)
50. Wan, B., Zhou, D., Liu, Y., Li, R., He, X.: Pose-aware multi-level feature network for human object interaction detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9469–9478 (2019)
51. Wang, H., Zheng, W.s., Yingbiao, L.: Contextual heterogeneous graph network for human-object interaction detection. In: European Conference on Computer Vision. pp. 248–264. Springer (2020)
52. Wang, H., Ho, E.S.L., Shum, H.P.H., Zhu, Z.: Spatio-temporal manifold learning for human motions via long-horizon modeling. IEEE Transactions on Visualization and Computer Graphics **27**(1), 216–227 (2021). https://doi.org/10.1109/TVCG.2019.2936810
53. Wang, N., Zhu, G., Zhang, L., Shen, P., Li, H., Hua, C.: Spatio-temporal interaction graph parsing networks for human-object interaction recognition. In: Proceedings of the 29th ACM International Conference on Multimedia. pp. 4985–4993 (2021)
54. Wang, X., Girshick, R., Gupta, A., He, K.: Non-local neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7794–7803 (2018)
55. Xiu, Y., Li, J., Wang, H., Fang, Y., Lu, C.: Pose flow: Efficient online pose tracking. arXiv preprint arXiv:1802.00977 (2018)
56. Xu, B., Wong, Y., Li, J., Zhao, Q., Kankanhalli, M.S.: Learning to detect human-object interactions with knowledge. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2019)
57. Yan, S., Xiong, Y., Lin, D.: Spatial temporal graph convolutional networks for skeleton-based action recognition. In: Thirty-second AAAI conference on artificial intelligence (2018)
58. Zhang, P., Lan, C., Zeng, W., Xing, J., Xue, J., Zheng, N.: Semantics-guided neural networks for efficient skeleton-based human action recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1112–1121 (2020)
59. Zheng, S., Chen, S., Jin, Q.: Skeleton-based interactive graph network for human object interaction detection. In: 2020 IEEE International Conference on Multimedia and Expo (ICME). pp. 1–6. IEEE (2020)
60. Zhu, M., Ho, E.S.L., Shum, H.P.H.: A skeleton-aware graph convolutional network for human-object interaction detection. In: Proceedings of the 2022 IEEE International Conference on Systems, Man, and Cybernetics. SMC '22 (2022)
61. Zhuang, B., Wu, Q., Shen, C., Reid, I., Hengel, A.v.d.: Care about you: towards large-scale human-centric visual relationship detection. arXiv preprint arXiv:1705.09892 (2017)