IPGN: Interactiveness Proposal Graph Network for Human-Object Interaction Detection

Haoran Wang, *Graduate Student Member, IEEE*, Licheng Jiao, *Fellow, IEEE*, Fang Liu, *Senior Member, IEEE*, Lingling Li, *Member, IEEE*, Xu Liu, *Member, IEEE*, Deyi Ji, and Weihao Gan

***Abstract*— Human-Object Interaction (HOI) Detection is an important task to understand how humans interact with objects. Most of the existing works treat this task as an exhaustive triplet $\langle human, verb, object \rangle$ classification problem. In this paper, we decompose it and propose a novel two-stage graph model to learn the knowledge of interactiveness and interaction in one network, namely, Interactiveness Proposal Graph Network (IPGN). In the first stage, we design a fully connected graph for learning the interactiveness, which distinguishes whether a pair of human and object is interactive or not. Concretely, it generates the interactiveness features to encode high-level semantic interactiveness knowledge for each pair. The class-agnostic interactiveness is a more general and simpler objective, which can be used to provide reasonable proposals for the graph construction in the second stage. In the second stage, a sparsely connected graph is constructed with all interactive pairs selected by the first stage. Specifically, we use the interactiveness knowledge to guide the message passing. By contrast with the feature similarity, it explicitly represents the connections between the nodes. Benefiting from the valid graph reasoning, the node features are well encoded for interaction learning. Experiments show that the proposed method achieves state-of-the-art performance on both V-COCO and HICO-DET datasets.**

***Index Terms*— Human-object interaction detection, two-stage graph model, interactiveness proposal, interaction learning.**

Manuscript received March 23, 2021; revised July 6, 2021; accepted July 6, 2021. Date of publication July 16, 2021; date of current version July 22, 2021. This work was supported in part by the Key Scientific Technological Innovation Research Project by the Ministry of Education; in part by the State Key Program and the Foundation for Innovative Research Groups of the National Natural Science Foundation of China under Grant 61836009 and Grant 61621005; in part by the Key Research and Development Program in Shaanxi Province of China under Grant 2019ZDLGY03-06; in part by the Major Research Plan of the National Natural Science Foundation of China under Grant 91438201, Grant 91438103, and Grant 61801124; in part by the National Natural Science Foundation of China under Grant U1701267, Grant 62006177, Grant 61871310, Grant 61902298, Grant 61573267, and Grant 61906150; in part by the Fund for Foreign Scholars through the University Research and Teaching Programs 111 Project under Grant B07048; in part by the Program for Cheung Kong Scholars and Innovative Research Team in University under Grant IRT 15R53; in part by the ST Innovation Project from the Chinese Ministry of Education; in part by the National Science Basic Research Plan in Shaanxi Province of China under Grant 2019JQ-659; in part by the China Postdoctoral Fund under Grant 2019M663641 and Grant 2017M613081; in part by the Scientific Research Project of Education Department in Shaanxi Province of China under Grant 20JY023; in part by the Fundamental Research Funds for the Central Universities under Grant XJS201901, Grant XJS201903, Grant JBF201905, and Grant JB211908; and in part by the CAAI-Huawei MindSpore Open Fund. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Jiwen Lu. *(Corresponding author: Licheng Jiao.)*

Haoran Wang was with SenseTime, Beijing 100080, China. He is now with the Key Laboratory of Intelligent Perception and Image Understanding of Ministry of Education, International Research Center for Intelligent Perception and Computation, School of Artificial Intelligence, Xidian University, Xi'an 710071, China, and also with the Joint International Research Laboratory of Intelligent Perception and Computation, School of Artificial Intelligence, Xidian University, Xi'an 710071.

Licheng jiao, Fang Liu, Lingling Li, and Xu Liu are with the Key Laboratory of Intelligent Perception and Image Understanding of Ministry of Education, International Research Center for Intelligent Perception and Computation, School of Artificial Intelligence, Xidian University, Xi'an 710071, China, and also with the Joint International Research Laboratory of Intelligent Perception and Computation, School of Artificial Intelligence, Xidian University, Xi'an 710071, China (e-mail: lchjiao@mail.xidian.edu.cn).

Deyi Ji and Weihao Gan are with SenseTime, Beijing 100080, China.

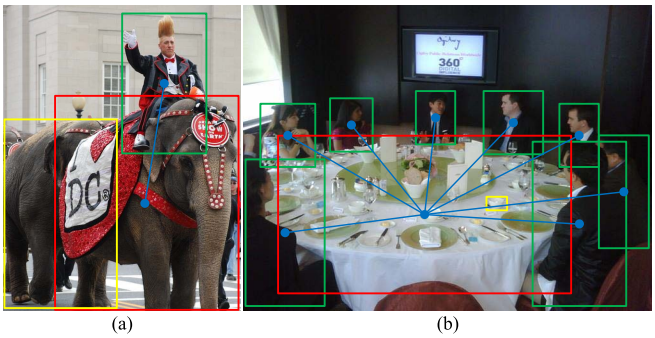
Digital Object Identifier 10.1109/TIP.2021.3096333

I. Introduction

HUMAN-OBJECT interaction (HOI) detection aims to localize and classify triplets of human, object and relation from an input image. Detecting the interaction can enable a well-designed algorithm to generate better descriptions for a scene. For instance, the image in Figure 1 (a) can be described as "The man is riding an elephant", rather than "There are a man and an elephant". The ability to predict such triplets and use them to understand human behavior in complex real-world environments is very valuable in activity analysis [1], socially-aware robots [2] and surveillance event detection [3].

In recent years, much progress has been made in scene understanding [1], [4]–[8], including object detection [9]–[13] and human pose estimation [14]–[16]. Driven by the impressive progress on these tasks, some deep learning methods have been proposed to address HOI detection problems. Generally, to tackle this task, one pretrained detector is used to detect humans and objects first. Then all humans and objects in the scene are paired exhaustively [17]. Finally, these pairs are classified as different HOI categories (including non-interaction category) based on various kinds of visual features. Through this pipeline, state-of-the-art works mainly focus on exploiting and fusing more informative visual features [18]–[22], and elaborate network design [23]–[29] to learning the objective. They directly make predictions for all the pairs based on the ability of **relation classification**. Despite their general efficacy, the existing pipeline is still insufficient for high-efficiency HOI solutions without considering **interactiveness**, which decides whether the action of the pair of HOI categories exists in the scene.

Directly classifying all the pairs based on the visual features lead to an ambiguity problem, there are too many negative pairs consisting of unrelated humans and objects. For example, as shown in the image in Figure 1 (b), the cup in the yellow box on the dining table is also detected with a high score. When paired with the present humans, the network needs to



(a) (b)

Fig. 1. Examples of riding an elephant (a) and sitting at the dining table (b) in HICO-DET dataset. Bounding boxes are detection results of the object detector. Green is for the human, red is for the object, yellow is for the negative object (without interactions) according to the ground truth. We only show one negative object in each image for concision.

decide whether the image contains an interactive triplet of $\langle human, drinkwith, cup\rangle$. Though the pair of $\langle human, cup\rangle$ indeed exists, it is a negative one without action in the task. The same scenario also happens to other objects like knife and tie, and we do not show them in the image for concision. By doing statistical analysis, there are more than 80% pairs that have no interactions with each other in most works, and removing these negative pairs can effectively improve the learning of interaction class decision. On the other hand, the HOI detection datasets suffer from long-tailed distributions. For the widely-used HICO-DET dataset [30], 138 out of the 600 HOI categories have fewer than 10 training samples. The problem is aggravated as the model need to distinguish negative pairs simultaneously.

Li *et al.* [26] first proposed to explore "Interactiveness Knowledge" which indicates whether human and object interact with each other or not. However, their method have two weaknesses. One is that it is not able to take full advantage of contextual information in the scene, as it processes each human-object pair independently, resulting in inadequate learning of relations. Second, the overall framework is not trained end-to-end, where the feature representations of interactiveness are not well utilized for HOI classification. Ulutan *et al.* [31] used an interaction score for this point. However, their method lacks the supervision and purposeful design, the interactiveness knowledge is not well learnt in the network. In conclusion, existing researches have not fully considered the interactiveness knowledge, which is implicit but important in the HOI detection task.

On the other hand, existing methods lacked distinctions when learning contextual information in the graph. For example, GPNN [17] connected all the humans and objects in a fully connected graph. Wang *et al.* [32] followed them and further proposed to regard humans and objects as different nodes. Meanwhile, some researches also provided two subgraphs where one connects the humans to objects and another connects the objects to humans [20], [31], [33]. All the methods mainly relied on two kinds of massage passing: appearance similarity and learnable parameters. But whatever the strategy they used, the transinformation happens between all the nodes in the graph, which may lead to redundant or even harmful connections.

To circumvent the issues, we propose a novel Interactiveness Proposal Graph Network (IPGN) for the HOI problem, which treats it as a two-stage classification task. Generally, the interactiveness classification is learnt in the first stage, and followed by the second interaction classification stage. In the first stage, we propose a fully connected graph where all pairs are distinguished whether they are interactive or not, providing reasonable proposals for the second stage. Moreover, the interactiveness knowledge is encoded into the interactiveness features which are also used in the second stage. Contrapose the previous works, the class-agnostic interactiveness is a more general and simpler objective, and the massage passing can be guided by explicit prior knowledge. Thus, we model the relations between nodes in the edge using both spatial layouts and features. In the second stage, a sparsely connected graph is constructed with all interactive pairs (proposals). Benefiting from the prior interactiveness knowledge, the network learns the contextual information among related nodes by valid message passing. Meanwhile the interactiveness features are used to update the node features in this stage, resulting in the effective interaction reasoning with the high-level semantic knowledge. The Graph Convolution Network (GCN) is directly used for the message passing in the graph. Finally, the output features in the second stage are paired and sent into the classifier for action prediction. Besides, by utilizing binary labels, we learn the interactiveness effectively under a strong supervision.

Overall, our contribution is threefold:

- We propose a novel Interactiveness Proposal Graph Network (IPGN) for HOI detection task. Especially, we propose to learn the interactiveness knowledge based on the fully connected graph. Under the binary supervision, it effectively reduces the negative pairs and alleviates the ambiguity problem. Moreover, the network is trained in an end-to-end way.
- We provide a sparsely connected graph based on the interactiveness proposals, which guarantees the validity of information flow among nodes. In this way, valuable contextual information can be learnt for interaction reasoning between humans or objects.
- Our method provides a unified framework for both interactiveness and interaction predictions. Experiments show that we achieve state-of-the-art results on both V-COCO and HICO-DET datasets.

II. Related Work

A. Visual Relationship Detection and Object Detection

Visual relationship detection [5], [7], [34], [35] aims to localize a pair of objects and recognize their relationships simultaneously. Lu *et al.* [5] proposed a relationship dataset VRD for this kind of task to give the predicates in a triplet $\langle human, predicate, object\rangle$. The predicates consist of actions, verbs, spatial and preposition vocabularies in their setting, which makes the task difficult as such vocabulary setting results in the long-tail issue in the task. The instrinsic

long-tail property of label space poses challenges in these tasks. To relieve the pressure, recent works [21], [36] put attention on the specific $\langle human, verb, object \rangle$ triplet classification, namely Human-object interaction detection.

For detecting HOIs, the first step is to detect humans and objects properly. With the significant progress of object detection, we are able to detect multi-scale objects in a image, which lays a foundation for HOI detection task. Generally, modern object detection approaches include RCNN [37], Faster RCNN [10], YOLO [38], SSD [39] and Feature Pyramid Network [11], which can be divided into single-stage and two-stage methods. Following the previous works, we choose to use a pre-trained Faster-RCNN model with the ResNet-50-FPN [11] backbone in our network to detect humans and objects.

B. Human-Object Interaction Detection

The task of human-object interaction detection aims to localize human, objects, as well as recognize the interactions between each pair of a person and an object. The earlier researches mainly focused on tackling HOI by utilizing human and object appearance features or their spatial layout [36], [40]. Recently, several methods [19], [20], [24], [25] have been developed to use more fine-grained visual features with advanced techniques, including human pose estimation [15] and attention mechanism [41]. In the pose-based method, Wan *et al.* [19] utilized human pose to capture global spatial configurations of relations and as an attention mechanism to zoom into relevant regions at the human part level. Zhou *et al.* [20] used human pose estimation to construct an object-bodypart graph and a human-bodypart graph to capture relationships between body parts and surrounding instances. These methods are not favored since additional annotations and computation are indispensable, which bring a large workload and computational burden. For attention mechanism, iCAN [42] exploited an instance-centric attention module to highlight the information from object regions and facilitate HOI classification. Wang *et al.* [24] proposed a deep contextual attention framework to capture context-aware appearance features for human and object. PMFNet [19] also focused on pose-aware attention in human parts. Overall, methods employing attention mechanisms learn informative regions but treat each visual target (i.e., scene, human and object) separately, which are still insufficient to exploit interactive semantics.

C. Label Hierarchy in Multi-Label Learning

The hierarchical structure of label categories has long been exploited for multi-label learning in various vision tasks, e.g., image/object classification [43], [44], detection [45], [46], and human pose estimation [47], [48]. By contrast, label hierarchy has rarely been considered in HOI detection. Kim *et al.* presented the first method to take advantage of an action label hierarchy for HOI recognition. They used co-occurring actions as prior knowledge and defined by co-occurrences rather than semantics or a taxonomy. Inspired by the previous works, we propose another way to decouple the HOI detection task into a two-level classification task: interactiveness and interaction recognition. We first judge whether a human-object pair has relations, and abandon the pairs that have no interactions, which is a binary classification task. Secondly, we predict the exact relation between the left pairs, recognizing the right verb category of the action, which is a multiclass classification task.

D. Graph Neural Networks

Some approaches have been proposed that apply deep neural networks to graph-structured data [49]–[55]. Kipf *et al.* [49] proposed a simple and well-behaved layer-wise propagation rule for neural network models and demonstrated its effectiveness in semi-supervised classification tasks. Tang *et al.* [54] proposed to use a "self-learned" weight matrix to transfer relation knowledge across two graphs under the guidance of a task-specific loss function. Li *et al.* [55] constructed a star-shaped graph to reason about the relations between two facial image features, which was further improved by a hierarchical reasoning graph network. These years, graph neural networks have been proposed to explore contextual information in the HOI detection task, as HOI is a natural graph where human and objects are connected. Qi *et al.* [17] proposed GPNN incorporating DNN and graph model, which uses message parsing to update states iteratively and classifies all possible human-object pairs. It was the first attempt to introduce a graph model structure for this task. RPNN [20] proposed Object-Bodypart Graph and Human-Bodypart Graph to capture the relationship between body parts and the surroundings, and assembled body part contexts to predict actions. DRG [33] proposed a human-centric subgraph and an object-centric subgraph based on the feature of spatial layout and word embedding. VSGNet [31] improved the idea of GPNN and updated the values of edges by learning model instead of visual similarity, simultaneously changed the fully connected graph into two subgraphs. Our method differs from them in three points: First, we propose a fully connected graph and a sparsely connected graph to learn the knowledge of interactiveness and interaction simultaneously. Second, we explicitly analyze the effect of different kinds of features for graph construction and encode the human and object features to a unified semantic space for graph reasoning. Third, we incorporate the interactiveness knowledge to update the node features, guiding the message passing with the high-level semantic knowledge in the sparse graph.

III. METHOD

In this section, we present the IPGN for HOI detection (Figure 2). We first give the problem formulation (Section III-A), then we show the detail of the overview of our network (Section III-B). Finally, we introduce basic branch (Section III-C) and graph branch (Section III-D) of our network.

A. Problem Formulation

Formally, given a set of human and object bounding boxes b_h, b_o and corresponding detection scores s_h, s_o, the score $S^a_{h,o}$ is required for each $< b_h, b_o >$ pair representing the

probability of actions. In our work, we define the score $S^a_{h,o}$ of each pair as:

$$S^a_{h,o} = s_h \cdot s_o \cdot (s^a_h \cdot s^a_o \cdot s^a_{h,o;spatial} \cdot s^a_{h,o;graph}) \cdot \dagger(s^{in}_{h,o} \geq \mu_s) \tag{1}$$

where s^a_h/s^a_o is the action prediction score from human/object representation. $s^a_{h,o;spatial}$ and $s^a_{h,o;graph}$ are the action prediction scores of the human-object pairwise representation from basic and graph branch respectively. $s^{in}_{h,o}$ is the interactiveness prediction score of the human-object pair from graph branch. $\dagger(\cdot)$ is the indicator function and μ_s acts as a score threshold which is a hyper-parameter.

B. Overview

In HOI detection, human and object need to be detected first. In this work, we follow the setting of [33] and employ Faster R-CNN [10] to prepare bounding boxes B and detection scores S. Then we filter detection results by the detection score thresholds. Obtaining the detected bounding boxes, we first extract the ROI pooled features for every human and object, which are viewed as appearance features in this work. Our network consists of two branches, including a basic branch and a graph branch (Figure 2). The basic branch consists of three streams for the action prediction based on the human, object and spatial features individually, which is shown in Figure 2 for details. Figure 3 illustrates the pipeline of the graph branch. Apart from the spatial coordinate and appearance feature, the word embedding of each object is also used, which is produced by fastText [56]. In this branch, we first propose a semantic encoding module to fuse three kinds of features into a unified space. Then we design two stages to predict the interactiveness and interaction, respectively. The first stage introduces a fully connected graph with all instances in the scene. Under the binary interactiveness supervision, the network learns to distinguish all the interactive pairs from negative ones, generating proper proposals. Therefore, the second stage can build a sparse graph where the nodes are the interactive humans and objects. Furthermore, the message passing and updating include the interactiveness features from the first stage. In this way, the contextual information is well explored by the proper information flow among related nodes, resulting in the effective interaction reasoning. Specifically, benefiting from the first stage, there is no need to distinguish human nodes from object nodes in the second stage. As we learn the interactiveness classification and use it to guide the connection in the second stage directly. Moreover, a higher value represents a stronger interactiveness of a pair, thus the confidence for the edge weight between this two nodes is greater. Lastly, we fuse the action scores from the two branches to produce the final predictions. Especially, we add the interactiveness supervision in the graph branch, which explicitly discriminates the non-interactive pairs and suppresses them for HOI classification.

C. Basic Branch

As shown in Figure 2, the basic branch consists of three streams: human stream, object stream and spatial stream. Our human/object streams follow the standard HOI detection pipeline for feature extraction and classification. For each ROI pooled human/object feature, we pass it into a residual block followed by global average pooling and obtain the human appearance feature f_h and the object appearance feature f_o. We then apply a standard classification layer to obtain the action scores s^a_h (from human stream) and s^a_o (from object stream), the process can be defined as:

$$\begin{aligned} f_h &= GAP(Res(RoI(F, b_h))) \\ f_o &= GAP(Res(RoI(F, b_o))) \\ s^a_h &= \sigma(W_h \cdot f_h) \\ s^a_o &= \sigma(W_o \cdot f_o) \end{aligned} \tag{2}$$

where F represents ROI pooling feature maps, f_h and f_o are appearance feature of each human and object, W_h and W_o are the learning weights, $GAP(\cdot)$ is the global average pooling function, $\sigma(\cdot)$ is the sigmoid function.

The spatial stream aims to encode the spatial locations of instances. Its input is a two-channel binary spatial configuration map $F_{h,o}$ consisting of a human map and an object map. Human and object maps are all obtained from the detection results. In the human channel, the value is 1 in the human bounding box and 0 in other areas. The object channel is similar which has value 1 in the object bounding box and 0 elsewhere. In this work, three convolution layers followed by a GAP operation and a fully connected layer are adopted to analyze the binary tensor:

$$\begin{aligned} f_{h,o} &= GAP(Conv(F_{h,o})) \\ s^a_{h,o;spatial} &= \sigma(W_{h,o} \cdot f^{spatial}_{h,o}) \end{aligned} \tag{3}$$

where $f_{h,o}$ represents the spatial configuration of the human-object pair, $W_{h,o}$ is the learning weights, $\sigma(\cdot)$ is the sigmoid function similar to human/object streams. As the objects and humans are defined in different channels, using convolutions on the binary spatial configuration map allows the model to learn the possible spatial relations between humans and objects.

D. Graph Branch

As shown in Figure 3, we first propose to learn the interactiveness knowledge in a fully connected graph, where the nodes denote all the humans and objects, and the edge represents the relation between two nodes. Then we perform interactiveness reasoning by GCN under the supervision. In this way, the interactiveness recognition can be viewed as the binary classification of pair-wise nodes. Then we obtain the interactiveness value and use it to guide the connection in the second stage directly, thus we can put the human and object nodes in the same graph. Meanwhile a higher value represents a stronger interactiveness of a pair, and the confidence for the edge weight between this two nodes is greater.

1) Semantic Encoding Module: Previous graph-based works mainly used the binary spatial configuration map and appearance features as the node features, which are not in the same encoding space. In our method, we use the semantic module to encode them to a unified semantic space. Specifically,

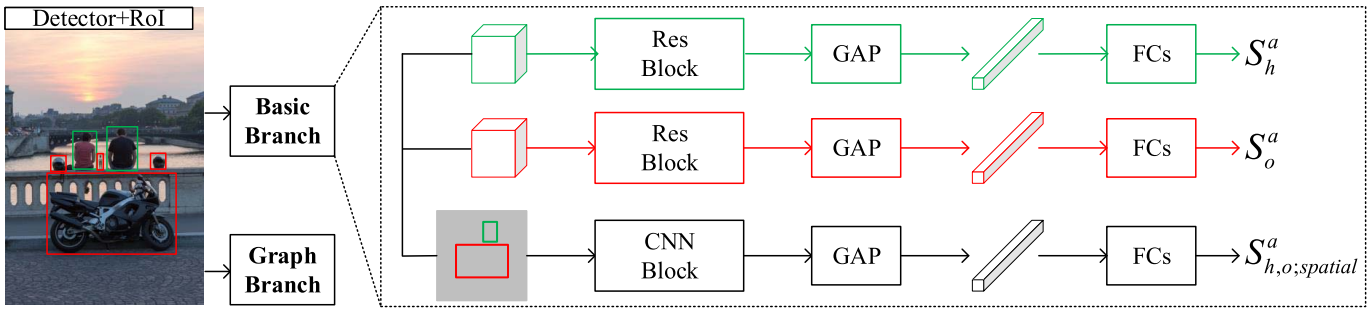


Fig. 2. Our IPGN includes the basic branch and the graph branch. The dotted box depicts the details of basic branch. The green, red and black represent human, object and spatial streams, respectively. The three streams provide the fundamental results for HOI classification without bells and whistles.

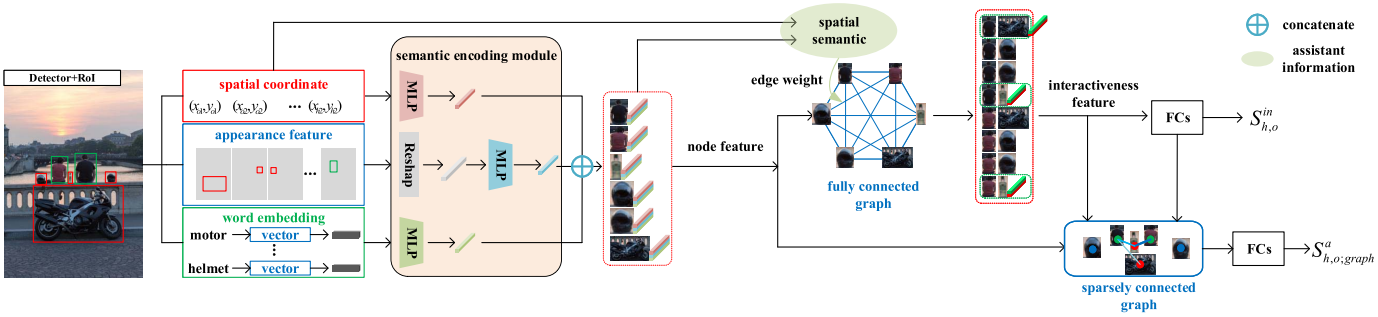


Fig. 3. The overview of the graph branch of IPGN network. Given an image, we first detect all the bounding boxes of humans and objects in the scene, then extract the ROI pooled appearance features for each box. Besides, the word embedding of each object is also obtained by fastText [56]. The box coordinates, appearance features and word embedding are used as the inputs of our network. In the first, we propose a semantic module to encode three features to a unified space for graph reasoning. In the first stage, we build a fully connected graph to learn the interactiveness for each pair under the interactiveness supervision, where the location and appearance information is explicitly considered. In the second stage, we build a sparsely connected graph to learn the interaction among related nodes under a novel message passing method, where the interactiveness features are incorporated. Finally, the updated features are used for the action classification, predicting the action score for each related pair.

we obtain the 2-d coordinates by the bounding box for each instance, then we take an MLP to encode the 2-d vector to 256-d spatial feature. Meanwhile, we use another MLP to encode the appearance feature for each instance. Besides, the word embedding of the object category is also taken into consideration, and we also use an MLP to encode this information for each instance. Finally, the three encoded features are concatenated in the channel dimension, resulting in a 768-d feature f for the graph node. We process all the humans and objects in the scene. The three kinds of features are useful cues when predicting the interactiveness. For example, it is very likely that two instances far apart are not interacting, and the objects of similar appearances and categories are perhaps related. We view them as the inherent characters for each instance and encode them to the same semantic space for the graph reasoning.

2) Fully Connected Graph: Formally, the fully connected graph corresponding to all instances in a scene is represented by an adjacent matrix: $A^f \in R^{N\times N} = \{(f_i)|i = 1, \ldots, N\}$, where N is the number of instances, f_i represents i-th encoded human/object feature. Recent advanced works formulate the adjacent matrix by learnable parameters because that general visual cues are not discriminative enough for too many action categories. However, visual factors explicitly reflect the interactiveness in our work. For example, the object may be not interactive without humans around; the human may be interactive with the cup as their spatial configuration has an overlap; the human perhaps rides a motor when wearing a helmet. Therefore, both spatial layouts and semantic correlations are considered. To this end, we model the spatial relation and semantic relation separately and explicitly, defined as below:

$$\begin{aligned} A_{ij}^f &= \frac{F_{dist}(i,j)exp(F_{se}(i,j))}{\sum_{j=1}^{N} F_{dist}(i,j)exp(F_{se}(i,j))} \\ F_{dist}(i,j) &= \frac{1}{D(b_i,b_j)} \\ F_{se}(i,j) &= (f_i)^T f_j \end{aligned} \tag{4}$$

where we perform softmax function on each node so that the sum of all the relation values of one node i will be 1. $F_{dist}(i,j)$ denotes the spatial relation between the two instances. $D(b_i,b_j)$ denotes the distance calculated by box coordinates of the two instances. $F_{se}(i,j)$ denotes the semantic relation between two nodes with the dot-product similarity. Moreover, graph-based methods mainly focus on designing various kinds of subgraphs where the human node only connects the object node and vice versa, it lacks the interactions and spatial configuration between human-human or object-object pairs. Actually, people often interact with each other in the real world. In our graph, all nodes are considered

in a union based on the same encoding space, the relations between these nodes are also learnt and contribute to the interactiveness prediction.

Taking the graph as input, the GCN performs reasoning over the structure. For a target node, it aggregates features from all neighbor nodes by the edge weight between them. Formally, one layer of the GCN can be written as:

$$g^{(l+1)} = \delta(Ag^l W^l) \tag{5}$$

where A represents the two kinds of adjacent matrixes. $g^l \in R^{N\times d}$ is the features of nodes in the lth layer, and $g^0 = f$ represents the encoded human/object feature. $W^l \in R^{d\times d}$ is the layer-specific learnable weight matrix, d is the size of the input and output feature. $\delta(\cdot)$ denotes an activation function, and we adopt ReLU in this work. The layer-wise propagation can be stacked into multi-layers, but we only use two layers of GCN for efficiency.

3) Sparsely Connected Graph: Previous works view the human and object as different nodes, where different message passing strategies [32] are usually needed to be designed (or using two subgraphs [31], [33] for a distinction). However, our method overcomes the issue in a sparse graph. On the one hand, all the graph nodes come from the interactive pairs (proposals) in the first stage. On the other hand, instead of utilizing the feature similarity between the nodes, we incorporate the class-agnostic interactiveness features to improve the message passing. The node features are then updated as a weighted addition fusion of all related nodes and the original features. Specifically, each node feature is also the encoded human/object feature, and the updating scheme of node feature f_i could be formulated as:

$$\begin{aligned} I^{in}_{i,j} &= \dagger(s^{in}_{h,o;i,j} \geq \mu_s) \cdot G^{in}_{i,j} \\ f^{(t)}_i &= (1-\alpha) f^{(t-1)}_i + \alpha \sum_{j=1}^{N} f^{(t-1)}_j I^{in}_{i,j} \end{aligned} \tag{6}$$

where $G^{in}_{i,j} = concat(g_i, g_j)$ denotes the interactiveness feature between the i-th and j-th nodes from the first stage, $\dagger(\cdot)$ is the indicator function, and μ_s acts as a score threshold. We learn to use it to select proper proposals which means the pairs have some specific interactions. $f^{(t)}_i$ denotes the i-th refined feature, $f^{(t-1)}_i$ denotes the i-th input feature, t is the iteration number, α represents the weighting parameter that balances fusion feature and original feature. The proposed feature fusion and updating are intuitive: using interactiveness scores to decide the connections between two nodes, meanwhile using the high-level semantic interactiveness features to update the node features.

The output features of two graphs are finally paired up and fed into two classifiers for interactiveness and HOI classification, respectively:

$$\begin{aligned} s^{in}_{h,o} &= \sigma(W_{in} \cdot f^{in}_{h,o}) \\ s^{a}_{h,o;graph} &= \sigma(W_a \cdot f^{graph}_{h,o}) \end{aligned} \tag{7}$$

where $f^{in}_{h,o}$ and $f^{graph}_{h,o}$ represent interactiveness features and interaction features of the human-object pairs, W_{in} and W_a are the learning weights corresponding with two classifiers, $\sigma(\cdot)$ is the sigmoid function. Obtaining all the scores, the action score $S^a_{h,o}$ is finally computed following the formula in Equation 1.

IV. Experiments

A. Experimental Setup

1) Dataset: We use the V-COCO [57] and HICO-DET datasets [30] to evaluate our model. V-COCO dataset is constructed by augmenting the COCO dataset [58] with additional human-object interaction annotations. It has 10,346 images. 2533 images are for training, 2867 images are for validating and 4946 images are for testing. Each person is annotated with a label of 29 different actions (five of them do not involve associated objects). If an object in the image is related to that action then the object is also annotated. HICO-DET is a larger dataset with 38118 training and 9658 testing images. It has 600 HOI categories over 80 object categories with more than 150K annotated instances of human-object pairs. Following [33], we select 117 object-agnostic action categories for training. At test time, we combine the predicted action and the detected object and convert them back to the original 600 HOI classes. The evaluation for the HICO-DET dataset remains the same.

2) Evaluation Metrics: The role mean average precision (role mAP) [57] is used to measure the performance for both V-COCO and HICO-DET datasets. The goal is to detect and correctly predict the triplet. We consider a prediction as true positive only if the human and object bounding boxes both have IoUs larger than 0.5 with reference to ground truth, and the HOI classification result is accurate.

3) Implementation Details: For fair comparison, we employ a Faster R-CNN [10] with the ResNet-50-FPN [11] backbone pretrained on MS-COCO as the object detector. A pretrained feature extractor with the ResNet-50 [59] backbone is adopted to extract image features. We keep the detected human boxes with scores higher than 0.7 and object boxes with scores higher than 0.3. We pair all the detected human and objects, and regard those who are not annotated in the ground-truth labels as negative examples. The ratio of positive and negative samples is 1:3. In the human/object stream, we take a residual convolution block for feature extraction, then a global average pooling and four 1024 sized fully-connected layers are used. In the spatial stream, we use a three-layer convolution block with a max-pooling layer and two 1024 sized fully-connected layers.

During training, we keep the ResNet-50 backbone frozen. Stochastic Gradient Descent (SGD) is used. The weight decay is 0.0001 and the momentum is 0.9. The batch size is set to 8. The initial learning rate is 0.001, and the cosine learning rate schedule is adopted. In the graph branch, the activation function after each layer of GCN is ReLU, and the dropout rate is 0.5. The threshold of interactiveness score is set to 0.5. The balancing weight α is set to 0.9. As each person can perform multiple actions on the same object simultaneously, we adopt binary sigmoid classifiers for multilabel classification. We train our model for 100 epochs. Our experiments are conducted on a single Nvidia Titan Xp GPU.

TABLE I

COMPARISONS WITH THE STATE-OF-THE-ART METHODS ON THE V-COCO TEST SET. THE FIRST PARTITION INCLUDES THE GRAPH-BASED MODELS, AND OUR NETWORK BELONGS TO THIS KIND OF METHOD. THE HIGHEST SCORES ARE BOLD. THE CHARACTER $*$ REPRESENTS THAT THE METHOD USES TWO SUBGRAPHS, AND ‡ REPRESENTS THAT THE METHOD USES MORE VISUAL OR SEMANTIC FEATURES WITH ADDITIONAL TECHNIQUES

Method	Feature backbone	mAP_{role}
GPNN [17]	ResNet-101	44.0
RPNN$*$‡ [20]	ResNet-50	47.5
In-GraphNet [60]	ResNet-50	48.9
DRG$*$ [33]	ResNet-50-FPN	51.0
VSGNet$*$ [31]	Resnet-152	51.8
Wang et al. [32]	ResNet-50-FPN	52.7
InteractNet [36]	ResNet-50-FPN	40.0
iCAN [42]	ResNet-50	45.3
Wang et al. [27]	Hourglass-104	51.0
PMFNet ‡ [19]	ResNet-50-FPN	52.0
ACP ‡ [61]	ResNet-152	53.0
FCMNet ‡ [62]	ResNet-50	53.1
Baseline	ResNet-50-FPN	47.3
IPGN	ResNet-50-FPN	**53.79**

TABLE II

COMPARISONS WITH THE STATE-OF-THE-ART METHODS ON THE HICO-DET TEST SET. THE HIGHEST SCORES ARE BOLD, AND THE CAPTIONS FOR THE PARTITION AND CHARACTERS ARE THE SAME AS THE ABOVE TABLE

Method	Feature Backbone	Default		
		Full	Rare	Non-Rare
GPNN [17]	ResNet-101	13.11	9.34	9.34
RPNN$*$‡ [20]	ResNet-50	17.35	12.78	18.71
Wang et al. [32]	ResNet-50-FPN	17.57	16.85	17.78
In-GraphNet [60]	ResNet-50	17.72	12.93	19.31
DRG$*$ [33]	ResNet-50-FPN	19.26	17.74	19.71
VSGNet$*$ [31]	Resnet-152	19.80	16.05	20.91
HO-RCNN [30]	CaffeNet	7.81	5.37	8.54
InteractNet [36]	ResNet-50-FPN	9.94	7.16	10.77
iCAN [42]	ResNet-50	14.84	10.45	16.15
PMFNet ‡ [19]	ResNet-50-FPN	17.46	15.65	18.00
Wang et al. [27]	Hourglass-104	19.56	12.79	21.58
FCMNet ‡ [62]	ResNet-50	20.41	17.34	21.56
ACP ‡ [61]	ResNet-152	20.59	15.92	21.98
Baseline	ResNet-50-FPN	17.08	13.34	18.16
IPGN	ResNet-50-FPN	**21.26**	**18.47**	**22.07**

During inference, we multiply all the prediction scores from two branches of our network. Additionally, we also multiply the final score with the detection scores from the object detector. Particularly, with a convincing prediction for interactiveness, we take the indicator function to decrease the loss from too many negative pairs.

B. Comparisons With the State of the Art

To validate the effectiveness, we compare with recent state-of-the-art methods on both of the datasets. The HOI detection result is evaluated with mean average precision following the settings provided by [57] and [30] for V-COCO and HICO-DET, respectively.

Table I shows the results on V-COCO dataset. The baseline model achieves the mAP of 47.3%, and our method improves it to the mAP of 53.79%, which outperforms all the existing models. Table II shows the results on the HICO-DET dataset in the setting of Default. Our method also provides competitive performance with state-of-the-art methods, achieving 21.26%, 18.47% and 22.07% on the *full*,*rare* and *non-rare* categories, respectively. Graph structure is appropriate to denote the HOI task in the scene, and we particularly compare the methods that are graph-based models [17], [20], [31]–[33], [60] in the first part of Table I and Table II. Some of them used sub-graphs to obtain the contextual information between heterogeneous entities (human-object), which neglects the influence of the homogeneous entities (human-human and object-object). In this case, VSGNet achieved the best performance of 51.8% on the V-COCO dataset with the Resnet-152 backbone. Lately, Wang *et al.* [32] proposed to also study heterogeneous entities (human-object) in one graph, bring a large improvement for the V-COCO dataset with 52.7% mAP. However, they failed to perform well on the HICO-DET dataset with only 17.57% mAP on the *full* categories. We assume that a proper prior sometimes servers better than learnable parameters in generalization. Different from them, we propose to utilize the explicit visual cues to construct the relation edges, and introduce the interactiveness supervision in the fully connected graph. Moreover, we utilize the interactiveness features in the sparse graph for valid node updating. In this way, we perform the best on both datasets and show the success in learning interactiveness knowledge that can further contribute to the interaction reasoning.

We also list two recent works that achieved the closest results to our work: FCMNet [62] and ACP [61]. Actually, FCMNet has used more visual features in their work. Though the human pose estimation is abandoned, they used human parsing and object segmentation masks for fine-grained features. Furthermore, they used the technique of predicting the optical flow to obtain motion features. Similarly, ACP used human pose based on a better backbone. We believe that our performance can also be improved with more features or better backbone, but this work focuses on implementing the HOI task with less computation effort. Especially, the interactiveness learning helps to reduce the negative pairs effectively. Table II shows that we increase the mAP by 5.13% on the *rare* categories compared with the baseline model, proving the ability to handle the long-tailed distributions of HOI classes of our method. In conclusion, we use less effort to achieve the highest result, providing an empirical two-stage framework for future researches.

C. Ablation Studies

1) Analysis of Individual Streams and Branches: Our overall architecture consists of the graph and basic branches. To analyze the effectiveness of each part, we evaluate the performance of them on two datasets individually in Table III and Table IV. More concretely, the basic branch consists of three streams as shown in Figure 2. The human and object streams consider two kinds of instances individually, they often provide poor performance without using contextual information. In our work, we build them and achieve the mAP of 46.61% and 45.59% on the V-COCO dataset. Following Li *et al.* [26], we re-implement the spatial stream and get the

TABLE III

PERFORMANCE OF THE BASIC BRANCH AND EACH STREAM ON THE V-COCO AND HICO-DET (*Full*) TEST SET

Method	V-COCO	HICO-DET
basic branch	**47.30**	**17.08**
human steam	46.61	15.84
object stream	45.59	14.26
spatial stream	46.83	17.62

TABLE IV

PERFORMANCE OF EACH PART IN GRAPH BRANCH ON THE V-COCO AND HICO-DET (*Full*) TEST SET

Method	V-COCO	HICO-DET
graph branch	**52.93**	**20.37**
w/o semantic module	50.36	18.41
w/o relation edges	48.21	17.16
w/o interactiveness supervision	49.25	17.51
w/o sparsely connected graph	47.15	16.23
w/o interactiveness feature	48.91	17.35

TABLE V

PERFORMANCE COMPARISONS OF THREE FEATURES. BASIC BRANCH IS USED, AND THE FEATURES ONLY CHANGE IN THE GRAPH BRANCH

Method	spatial coordinate	appearance feature	word embedding	mAP
Ours				47.30
	✓			50.21
		✓		51.47
			✓	50.63
	✓	✓		52.31
	✓		✓	51.98
		✓	✓	52.47
	✓	✓	✓	**53.79**

TABLE VI

COMPARISONS WITH EXISTING METHODS ON THE REDUCTION OF NON-INTERACTIVE PAIRS AND CORRESPONDING RESULTS ON TWO DATASETS. BOTH ARE TRAINED AND TESTED ON THE SAME DATASET

Reduction (%) / mAP (%)	V-COCO	HICO-DET
RP_DC_D [26]	-65.98 / 47.8	-65.96 / 17.03
IPGN	-71.21 / **53.79**	-70.37 / **21.26**

result of 46.83% mAP. These streams provide fundamental results like previous works.

In Table IV, we further analyze the details of the graph branch. Generally, our graph branch already obtains competitive performance compared to the state-of-the-art with the highest mAP of 52.93%. The semantic encoding module encodes the three features of humans and objects to a unified space. Instead, we directly concatenate three original features for each instance to obtain the node feature. The performance decrease approximately 2% on both datasets. The relation edges guarantee the validity of message passing between the nodes, based on the appropriate priors of spatial and semantic relations. To verify the effectiveness of this method, we view all the neighbourhood with the same weight of 1. In this condition, the performance drops 4.72% and 3.21% on the two datasets, which may be caused by redundant or even harmful connections between nodes without guidance. Finally, we take out of the interactiveness supervision, and get the 3.68% and 2.86% decrease in mAP. Actually, these parts are complementary to each other in the graph, the relation graph may not perform well without the general supervision of the interactiveness. For the setting of w/o sparsely connected graph, we directly build one-stage fully connected graph model to predict the HOI actions. It shows a large drop in accuracy on both datasets, which are 47.15% and 16.23% mAP, respectively. For the setting of w/o interactiveness feature, the message passing in the sparsely connected graph is the simple fusion of only original features. In this case, the interactiveness knowledge is not used by the graph reasoning for interactions, which also shows a decrease of 4.02% and 3.02% mAP on the two datasets.

2) Efficacy of Multi-Modal Features: Table V reports the efficacy of three features on V-COCO test set. The first row shows the baseline result of 47.30% mAP. Using the bounding boxes predicted by detector, we generate the spatial coordinates and achieve the mAP of 50.21%. As the coordinates can be obtained directly without other techniques, it shows impressive improvement of 2.91% in our method. Then we add the appearance feature and word embedding separately. It shows that the appearance features bring the greatest benefit with an increase of 4.17%. And the word embedding gives the increase of 3.33%. Then we perform the combinations of every two features, and the we get similar performance gain as the single one. The combination of spatial coordinates and word embedding get the minimum increase with 51.98% mAP. When combining the appearance feature and word embedding, we obtain the highest score of 52.47%. Finally, the mAP can be improved to 53.79% when the three features are used together. The performance of single feature is relatively lower than the combination as the complementary role of multi-modal features.

3) Non-Interaction Reduction: Li *et al.* [26] also proposed to learn the interactiveness knowledge in an individual network to remove the negative pairs, and we show the comparisons with them in Table VI. It shows that our method gives a reduction of 71.21% and 70.37% in quantity on the two datasets, and their method gave 65.98% and 65.96% respectively. They used the ResNet-50 as the backbone of the object detector, but the reduction(%) is measured under fair conditions, which is a relative result of our and their methods. As a result, they only achieved 47.8% and 17.03% mAP on two datasets, which are much lower than ours. By contrast, we train the model in an end-to-end way and integrate interactiveness knowledge into the interaction learning, showing a superior framework than their method. Furthermore, we visualize the reduction behaviour under the help of interactiveness recognition in Figure 4. Experimentally, the scene full of the messy objects is more likely to have non-interactive pairs, such as a dining table. However, our method still performs well in such scenarios. Besides, sometimes the non-interactive pairs still provide the correct HOI predictions as demonstrated in the last image. Though the human has no interactions with another bicycle, they have the predicted class of "hold/sit on bicycle" when paired in a group. The phenomenon provides evidence that the previous works could perform not badly where the whole scene is strongly correlated. However, this may lay a hidden danger in the future practical application.

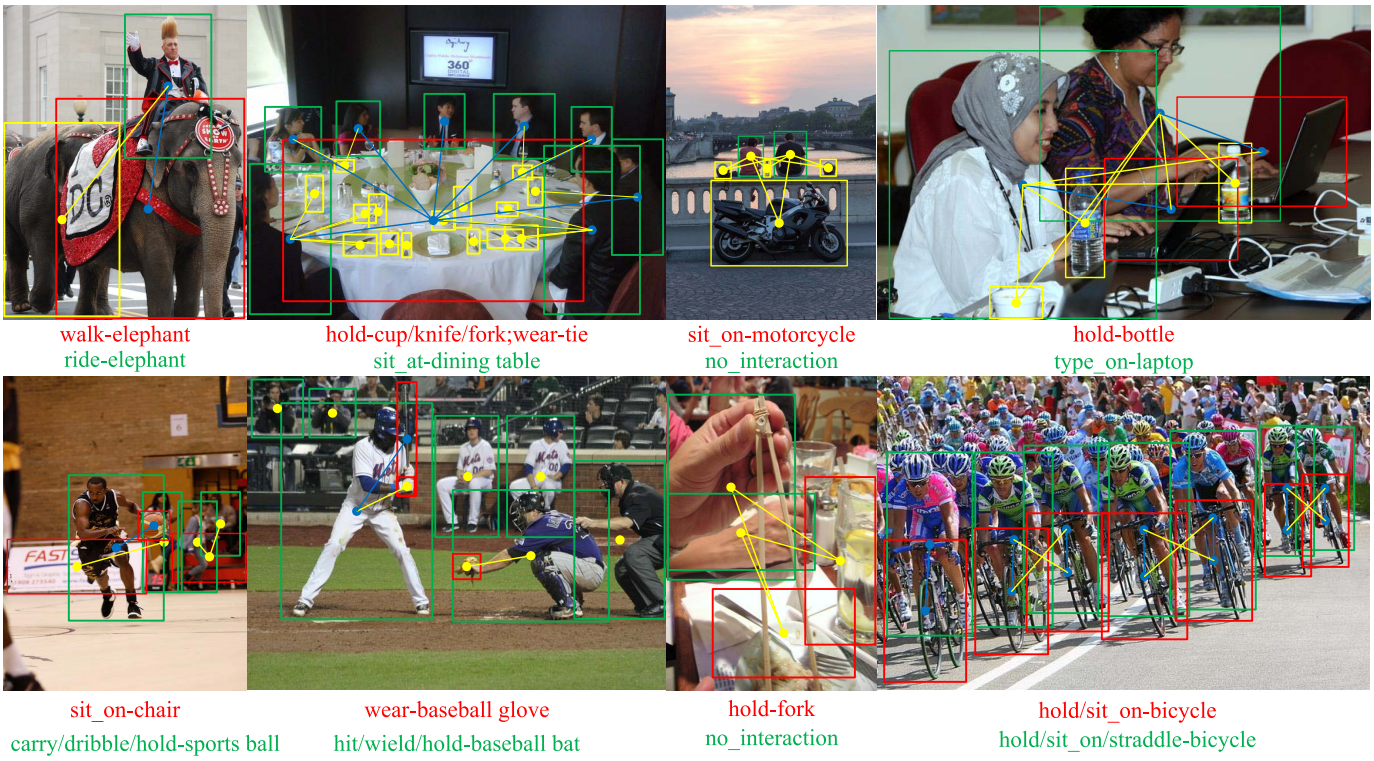


Fig. 4. Visualization of the reductions under the help of interactiveness recognition. Blue lines mean accurate pairs, yellow lines mean non-interactive pairs which are removed. For better understanding the ambiguity problem, we give the false positive predictions for the non-interactive pairs in red texts, and true positive predictions for the interactive pairs in green texts, which are shown below each image.

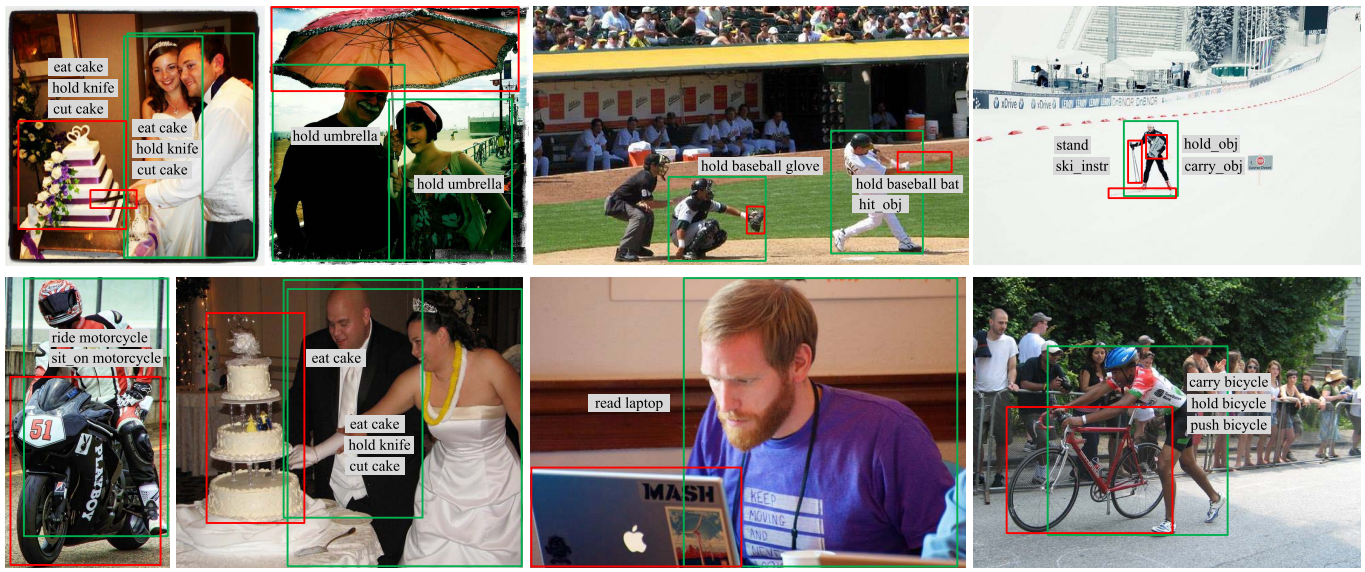


Fig. 5. Qualitative results of HOI detection on V-COCO and HICO-DET test set, which are shown in the first and second rows, respectively. Human and object instances are denoted by green and red bounding boxes.

4) Comparisons of Per-Class Performances: We report per-class performances in Table VII on the V-COCO dataset. Compared with the existing methods which reported the values, we achieve better performance in major classes. With the help of the prediction in interactiveness, we perform well in some crowded scene where many objects exist but are noninteractive. For example, the class of sit-instr is often related to a table, and there are often many cups and bottles on it. In our network, such objects which have no interactions with humans are abandoned, eliminating the possibility of correlative classes of hold-obj or drink-instr. Besides, same as VSGNet, some of the action classes perform poorly due to the failure of object detectors like eat-instr, and we believe this can be improved by better detectors.

5) Qualitative Results: Here we supply more results on the V-COCO dataset and the HICO-DET dataset in Figure 5. The first row shows the samples of V-COCO dataset and the second row shows the samples of HICO-DET dataset. The person

TABLE VII

PER CLASS AP COMPARISONS TO THE EXISTING METHODS ON THE V-COCO DATASET. WE ONLY COMPARE TO THE METHODS WHICH HAVE REPORTED THE PER CLASS AP VALUES. OBJ REFERS OBJECT AND INSTR REFERS TO THE INSTRUMENT. THE HIGHEST SCORES ARE BOLD IN EACH COMPARISON. NOTE THAT VSGNET USED A BETTER RESNET-152 BACKBONE

HOI Class	InteractNet [36]	iCAN [42]	VSGNet [31]	IPGN
hold-obj	26.38	29.06	48.27	**51.12**
sit-instr	19.88	26.04	29.9	**31.82**
ride-instr	55.23	61.9	**70.84**	69.72
look-obj	20.2	26.49	42.78	**45.28**
hit-instr	62.32	74.11	**76.08**	75.6
hit-obj	43.32	46.13	48.6	**53.51**
eat-obj	32.37	37.73	38.3	**41.77**
eat-instr	1.97	**8.26**	6.3	7.9
jump-instr	45.14	51.45	52.66	**55.11**
lay-instr	20.99	22.4	21.66	**26.27**
talk on phone	31.77	52.81	62.23	**66.33**
carry-obj	33.11	32.02	39.09	**44.76**
throw-obj	40.44	40.62	45.12	**50.38**
catch-obj	42.52	47.61	44.84	**48.87**
cut-instr	22.97	37.18	**46.78**	45.69
cut-obj	36.4	34.76	36.58	**40.27**
work on comp	57.26	56.29	**64.6**	64.5
ski-instr	36.47	41.69	50.59	**54.34**
surf-instr	65.59	77.15	**82.22**	81.3
skateboard-instr	75.51	79.35	**87.8**	86.6
drink-instr	33.81	32.19	54.41	**56.27**
kick-obj	69.44	66.89	**69.85**	68.96
read-obj	23.85	30.74	42.83	**45.19**
snowboard-instr	63.85	74.35	**79.9**	79.5
Average	40.0	45.3	51.76	**53.79**

and the associated object are highlighted with green and red bounding boxes. respectively. In the cases where two persons have interactions with a cake, our method distinguishes the subtle visual differences between the man and the woman. For example, in the first image of V-COCO dataset, both of them hold the knife and cut the cake. By contrast, in the second image of HICO-DET dataset, the man has no interactions with the knife, and our method succeeds to avoid the predictions of holding a knife and cutting a cake. Meanwhile, it shows the ability to capture multiple actions based on various visual cues, which indicates that our model can detect co-occurrence relations. In the third image of V-COCO dataset, the right two man are detected with actions such as holding a baseball glove or hitting an object. However, the left person has no interactions with any object in the baseball field. It shows that our model succeeds to remove the one to alleviate the effects of negative samples, which is benefitted from the interactiveness learning in the first stage in our method. Moreover, many background people can be easily neglected with a lower detection score. As a result, the potential ambiguity problem can be better addressed by our network framework. There are also many examples to prove the effectiveness of our method, which improves the performance on both datasets.

V. CONCLUSION

In this paper, we present a novel Interactiveness Proposal Graph network for human-object interaction detection. Specifically, we decouple the task into two aspects of learning for interactiveness and interaction. More concretely, we propose to learn the interactiveness knowledge in a fully connected graph, which is well constructed by the explicit relation edges under the more general and more superficial objective. To learn the interaction knowledge, a novel graph reasoning is carried out in a sparse graph with the integration of interactiveness knowledge. Specifically, the proposed first-stage learning provides the proposals that guarantees the existence of relations, thus there is no need to distinguish human nodes from object nodes in the second stage. In this case, a higher value represents a stronger interactiveness of a human-object pair. Meanwhile the interactiveness features are used to update the node features in the second stage, resulting in the effective interaction reasoning with the high-level semantic knowledge. To validate the effectiveness of our method, we conduct all experiments on two widely used HOI detection datasets: V-COCO and HICO-DET. Experiments show that our method outperforms the state-of-the-art method on the two datasets, proving the effectiveness and superiority of our method. In the future, we expect a broader range of applications based on our method, especially for the relation recognition tasks that can utilize the graph model to decouple themselves into two-stage classifications.

REFERENCES

[1] F. C. Heilbron, V. Escorcia, B. Ghanem, and J. C. Niebles, "ActivityNet: A large-scale video benchmark for human activity understanding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 961–970.

[2] J. K. Westlund *et al.*, "Tega: A social robot," in *Proc. 11th ACM/IEEE Int. Conf. Hum.-Robot Interact. (HRI)*, Mar. 2016, p. 561.

[3] A. Adam, E. Rivlin, I. Shimshoni, and D. Reinitz, "Robust real-time unusual event detection using multiple fixed-location monitors," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 3, pp. 555–560, Mar. 2008.

[4] G. Wang, A. Gallagher, J. Luo, and D. Forsyth, "Seeing people in social context: Recognizing people and social relationships," in *Proc. Eur. Conf. Comput. Vis.* Berlin, Germany: Springer, 2010, pp. 169–182.

[5] C. Lu, R. Krishna, M. Bernstein, and L. Fei-Fei, "Visual relationship detection with language priors," in *Proc. Eur. Conf. Comput. Vis.* Berlin, Germany: Springer, 2016, pp. 852–869.

[6] Y. Zhang, L. Cheng, J. Wu, J. Cai, M. N. Do, and J. Lu, "Action recognition in still images with minimum annotation efforts," *IEEE Trans. Image Process.*, vol. 25, no. 11, pp. 5479–5490, Nov. 2016.

[7] R. Krishna *et al.*, "Visual genome: Connecting language and vision using crowdsourced dense image annotations," *Int. J. Comput. Vis.*, vol. 123, no. 1, pp. 32–73, 2017.

[8] H.-S. Fang, G. Lu, X. Fang, J. Xie, Y.-W. Tai, and C. Lu, "Weakly and semi supervised human body part parsing via pose-guided knowledge transfer," 2018, *arXiv:1805.04310*. [Online]. Available: http://arxiv.org/abs/1805.04310

[9] Y. T. Chen and C. S. Chen, "Fast human detection using a novel boosted cascading structure with meta stages," *IEEE Trans. Image Process.*, vol. 17, no. 8, pp. 1452–1464, Aug. 2008.

[10] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 91–99.

[11] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2117–2125.

[12] G. Cheng, J. Han, P. Zhou, and D. Xu, "Learning rotation-invariant and Fisher discriminative convolutional neural networks for object detection," *IEEE Trans. Image Process.*, vol. 28, no. 1, pp. 265–278, Jan. 2019.

[13] S. Zhou *et al.*, "Hierarchical and interactive refinement network for edge-preserving salient object detection," *IEEE Trans. Image Process.*, vol. 30, pp. 1–14, 2021.

[14] M. Ding and G. Fan, "Articulated and generalized Gaussian kernel correlation for human pose estimation," *IEEE Trans. Image Process.*, vol. 25, no. 2, pp. 776–789, Feb. 2016.

[15] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2D pose estimation using part affinity fields," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 7291–7299.

[16] X. Chu, W. Yang, W. Ouyang, C. Ma, A. L. Yuille, and X. Wang, "Multi-context attention for human pose estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1831–1840.

[17] S. Qi, W. Wang, B. Jia, J. Shen, and S.-C. Zhu, "Learning human-object interactions by graph parsing neural networks," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 401–417.

[18] P. Wei, Y. Zhao, N. Zheng, and S.-C. Zhu, "Modeling 4D human-object interactions for joint event segmentation, recognition, and object localization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1165–1179, Jun. 2017.

[19] B. Wan, D. Zhou, Y. Liu, R. Li, and X. He, "Pose-aware multi-level feature network for human object interaction detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9469–9478.

[20] P. Zhou and M. Chi, "Relation parsing neural network for human-object interaction detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 843–851.

[21] B. Xu, Y. Wong, J. Li, Q. Zhao, and M. S. Kankanhalli, "Learning to detect human-object interactions with knowledge," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1–10.

[22] Y.-L. Li *et al.*, "PaStaNet: Toward human activity knowledge engine," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 382–391.

[23] B. Yao and L. Fei-Fei, "Recognizing human-object interactions in still images by modeling the mutual context of objects and human poses," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 9, pp. 1691–1703, Sep. 2012.

[24] T. Wang *et al.*, "Deep contextual attention for human-object interaction detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 5694–5702.

[25] T. Xiao, Q. Fan, D. Gutfreund, M. Monfort, A. Oliva, and B. Zhou, "Reasoning about human-object interactions through dual attention networks," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 3919–3928.

[26] Y.-L. Li *et al.*, "Transferable interactiveness knowledge for human-object interaction detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3585–3594.

[27] T. Wang, T. Yang, M. Danelljan, F. S. Khan, X. Zhang, and J. Sun, "Learning human-object interaction detection using interaction points," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 4116–4125.

[28] Z. Ji, X. Liu, Y. Pang, W. Ouyang, and X. Li, "Few-shot human-object interaction recognition with semantic-guided attentive prototypes network," *IEEE Trans. Image Process.*, vol. 30, pp. 1648–1661, 2021.

[29] T. Zhou, S. Qi, W. Wang, J. Shen, and S.-C. Zhu, "Cascaded parsing of human-object interaction recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Jan. 5, 2021, doi: 10.1109/TPAMI.2021.3049156.

[30] Y.-W. Chao, Y. Liu, X. Liu, H. Zeng, and J. Deng, "Learning to detect human-object interactions," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2018, pp. 381–389.

[31] O. Ulutan, A. S. M. Iftekhar, and B. S. Manjunath, "VSGNet: Spatial attention network for detecting human object interactions using graph convolutions," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 13617–13626.

[32] S. Zheng, S. Chen, and Q. Jin, "Skeleton-based interactive graph network for human object interaction detection," in *Proc. IEEE Int. Conf. Multimedia Expo. (ICME)*, Jul. 2020, pp. 1–6.

[33] C. Gao, J. Xu, Y. Zou, and J.-B. Huang, "DRG: Dual relation graph for human-object interaction detection," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 696–712.

[34] M. A. Sadeghi and A. Farhadi, "Recognition using visual phrases," in *Proc. CVPR*, Jun. 2011, pp. 1745–1752.

[35] M. Yatskar, L. Zettlemoyer, and A. Farhadi, "Situation recognition: Visual semantic role labeling for image understanding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 5534–5542.

[36] G. Gkioxari, R. Girshick, P. Dollar, and K. He, "Detecting and recognizing human-object interactions," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8359–8367.

[37] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 580–587.

[38] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 779–788.

[39] W. Liu *et al.*, "SSD: Single shot multibox detector," in *Eur. Conf. Comput. Vis.* Berlin, Germany: Springer, 2016, pp. 21–37.

[40] T. Gupta, A. Schwing, and D. Hoiem, "No-frills human-object interaction detection: Factorization, layout encodings, and training techniques," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9677–9685.

[41] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," 2014, *arXiv:1409.0473*. [Online]. Available: http://arxiv.org/abs/1409.0473

[42] C. Gao, Y. Zou, and J.-B. Huang, "ICAN: Instance-centric attention network for human-object interaction detection," 2018, *arXiv:1808.10437*. [Online]. Available: http://arxiv.org/abs/1808.10437

[43] J. Deng *et al.*, "Large-scale object classification using label relation graphs," in *Proc. Eur. Conf. Comput. Vis.* Berlin, Germany: Springer, 2014, pp. 48–64.

[44] Z. Yan *et al.*, "HD-CNN: Hierarchical deep convolutional neural networks for large scale visual recognition," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 2740–2748.

[45] S. J. Hwang, F. Sha, and K. Grauman, "Sharing features between objects and their attributes," in *Proc. CVPR*, Jun. 2011, pp. 1761–1768.

[46] M. Marszalek and C. Schmid, "Semantic hierarchies for visual object recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2007, pp. 1–7.

[47] X. Sun, Y. Wei, S. Liang, X. Tang, and J. Sun, "Cascaded hand pose regression," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 824–832.

[48] X. Sun, C. Li, and S. Lin, "Explicit spatiotemporal joint relation learning for tracking human pose," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Oct. 2019.

[49] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," 2016, *arXiv:1609.02907*. [Online]. Available: http://arxiv.org/abs/1609.02907

[50] Y. Shen, H. Li, S. Yi, D. Chen, and X. Wang, "Person re-identification with deep similarity-guided graph neural network," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 486–504.

[51] J. Wu, L. Wang, L. Wang, J. Guo, and G. Wu, "Learning actor relation graphs for group activity recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 9964–9974.

[52] Y. Huang, H. Bi, Z. Li, T. Mao, and Z. Wang, "STGAT: Modeling spatial-temporal interactions for human trajectory prediction," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6272–6281.

[53] A. Mohamed, K. Qian, M. Elhoseiny, and C. Claudel, "Social-STGCNN: A social spatio-temporal graph convolutional neural network for human trajectory prediction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 14424–14432.

[54] Y. Tang, Y. Wei, X. Yu, J. Lu, and J. Zhou, "Graph interaction networks for relation transfer in human activity videos," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 9, pp. 2872–2886, Sep. 2020.

[55] W. Li, J. Lu, A. Wuerkaixi, J. Feng, and J. Zhou, "Reasoning graph networks for kinship verification: From star-shaped to hierarchical," *IEEE Trans. Image Process.*, vol. 30, pp. 4947–4961, 2021.

[56] T. Mikolov, E. Grave, P. Bojanowski, C. Puhrsch, and A. Joulin, "Advances in pre-training distributed word representations," 2017, *arXiv:1712.09405*. [Online]. Available: http://arxiv.org/abs/1712.09405

[57] S. Gupta and J. Malik, "Visual semantic role labeling," 2015, *arXiv:1505.04474*. [Online]. Available: http://arxiv.org/abs/1505.04474

[58] T.-Y. Lin *et al.*, "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.* Berlin, Germany: Springer, 2014, pp. 740–755.

[59] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[60] D. Yang and Y. Zou, "A graph-based interactive reasoning for human-object interaction detection," in *Proc. 29th Int. Joint Conf. Artif. Intell.*, Jul. 2020, pp. 1111–1117.

[61] D.-J. Kim, X. Sun, J. Choi, S. Lin, and I. S. Kweon, "Detecting human-object interactions with action co-occurrence priors," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 718–736.

[62] Y. Liu, Q. Chen, and A. Zisserman, "Amplifying key cues for human-object-interaction detection," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 248–265.