iCGPN: Interaction-Centric Graph Parsing Network for Human-Object Interaction Detection

Wenhao Yang, Guanyu Chen, Zhicheng Zhao, Fei Su and Hongying Meng

***Abstract*—Human-Object Interaction (HOI) detection aims to infer different interactions, which occur between humans and related objects of images. HOI is usually represented by a triplet $\langle human, action, object\rangle$ and can be modeled as a graph. Thus, with global structural information of images, graph-based methods can detect interactions. However, in existing graph networks, although different fully-connected graphs are built, all detected bounding boxes are regarded as graph nodes equally or different types of nodes according to the category, thereby the dominant role of humans in HOI is not obvious or ignored. In addition, object node representations mainly focus on appearance features, contributing little to HOI inference. To address these issues, a novel graph-based HOI detection model, named interaction-centric graph parsing network (iCGPN) is proposed here that models one human node as a central node, and other nodes as semantic nodes. Firstly, for each detected human bounding box, a human-centric fully-connected graph is constructed to learn related HOIs. Secondly, in order to reflect the difference between central nodes and semantic nodes and model different edge relationships, we design different feature representations. Through introducing an attention mechanism like a transformer, global information related to human-object interaction is explored to enrich the semantic node representation, in which spatial layout, relative locations and object categories information are combined. Finally, a multi-relation graph convolutional network is applied to update the node feature and infer the HOI. Furthermore, a hierarchical random shift is proposed to augment the data of the training set to fit the object detection deviation and enhance the network generalization ability. Extensive experimental results show that iCGPN achieves very competitive results in comparison with state-of-the-art graph-based methods on the V-COCO and HICO-DET datasets, which demonstrates the effectiveness of the proposed method.**

***Index Terms*—human-object interaction detection, attention mechanism, multi-relation graph convolutional network, hierarchical random shift.**

I. Introduction

In the past few years, object detection [5, 12, 15, 22, 34] has great progress due to the application of deep learning. However, relationship between objects are not sufficiently explored, thereby Human-Object Interaction (HOI) [8, 13, 17, 21] has been introduced to determine the relationship between humans and objects in given images, and many two-stage methods [7, 8, 21, 26, 28, 30, 33] were proposed, in which all human-object pairs are firstly enumerated, and then interactions of the pair are detected. Generally, HOI detection consists of two main steps: (1) object dete,ction and (2) HOI prediction. We firstly use an object detector to detect all human/object instances and then predict the interaction of all human-object pairs. HOI is usually represented by a triplet $\langle human, action, object\rangle$ as shown in Fig.1.

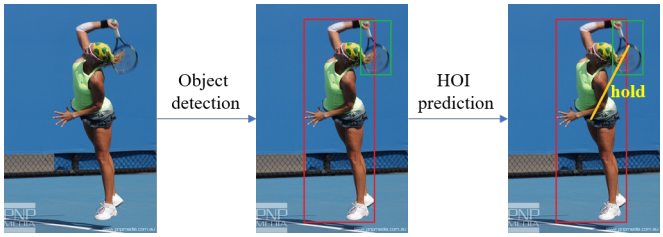

Fig. 1. HOI detection. Bounding boxes are detection results of the object detector. Green is for the human and red is for the object. The HOI result of this example is a triplet $\langle human, hold, tennis\ \ racket\rangle$.

Based on model architectures, the mainstream two-stage HOI detection can be divided into two groups: multi-stream methods [8, 9, 16, 19, 20, 26, 28] and graph-based methods [21, 27, 31, 33]. There are some other works [7, 18, 30, 32] combining them. Typical multi-stream methods contain three kinds of streams: human, object, and spatial stream. Recent attempts [19, 20] include the integrating of extra knowledge such as word embedding. The input of the multi-stream network is usually a human-object pair, which highlights the subjectivity of the human and can discover multiple information related to the HOI. Differently, graph-based methods make use of global structural knowledge by constructing graphs where detected boxes are treated as graph nodes, and then node representations are updated to predict interaction labels. In general, multi-stream methods can concentrate richer features in many aspects than graph-based methods as they are weaker in extracting relationship information. However, graph-based methods are easier to utilize context and global information and thus may get better performance in the relationship reasoning for a more comprehensive understanding. Anyway, several issues are still remaining.

Firstly, the dominant role of humans is not obvious, even ignored in most graph-based methods [21, 27]. We argue that in a real HOI graph, there should have only one central human node, and it is different from other nodes (may be objects and other human nodes). Therefore, we treat them as a central

This work is supported by Chinese National Natural Science Foundation (62076033, U1931202), and The National Key Research and Development Program of China (2020YFB2104600). (*Corresponding authors: Zhicheng Zhao*)

Wenhao Yang, Guanyu Chen, Zhicheng Zhao and Fei Su are with the School of Information and Communication Engineering and Beijing, Key Laboratory of Network System and Network Culture, Beijing University of Posts and Telecommunications, Beijing, China, (e-mail: whyang78@bupt.edu.cn; loraschen@bupt.edu.cn; zhaozc@bupt.edu.cn, sufei@bupt.edu.cn).

Dr.Hongying Meng is with the Department of Electronic and Electrical Engineering, Brunel University London, Uxbridge, United Kingdom, e-mail:hongying.meng@brunel.ac.uk.

Fig. 2. The relationship information among all object, such as spatial information and object category, can disambiguate in HOI detection. On the left of the picture, the model may mistakenly infer cut-with predicates for [human-knife] if it only depends on the visual and spatial feature of human and knife. But the relationship information of [sink-knife] is useful for the model to correctly infer wash predicates. On the right, with the message from [tennis racket-sports ball], the model can also easily infer hit predicates for [human-sports ball].

node and semantic nodes. Hence, different from Zheng, et al. [27], the relationship between human nodes and object nodes are different according to the role (central nodes or semantic nodes) of the human nodes. Hence, it is helpful to understand the scene and interaction (catching frisbee) between the central node and related three semantic nodes.

Secondly, some recent works [7, 18] neglect the relationship information among objects. In our opinion, the relationship among all objects obtained by visual feature, spatial information and object category, can disambiguate in HOI detection. For example, on the left side of Fig.2, the relationship of [sink-knife] is useful for correctly inferring wash instead of predicting cut-with for [human-knife]. On the right side, both two actions, i.e., throw or hit between human and ball may be predicted if only the visual and spatial cues are considered. Thus, by the supplement of contextual cue from [tennis racket-sports ball], the right action hit can be easily inferred. Therefore, a fully-connected graph can be constructed to build the relationships among all nodes.

Thirdly, unlike multi-stream methods, most graph models [18, 21, 27, 30–32] mainly use appearance features within object bounding boxes to infer HOIs, and richer features such as spatial and semantic features are neglected. Inspired by multi-stream methods, we aggregate visual, spatial and semantic information in parallel for HOI detection. Accordingly, for different types of nodes, different modules are designed to extract feature representations. Specially, via introducing an attention mechanism like a transformer [24], global information related to HOI is explored to enrich the semantic node representation. In addition, spatial layout, relative locations and object categories are also combined.

Moreover, during the inference, the detection model will be firstly applied to get ROIs, while the detection may not be good as that in training data. Considering that the IoUs between ground-truth and detected boxes ranges from 0.5 to 1, a hierarchical random shift method is proposed to augment the data of the training set, so as to relieve the inconsistency between the detected ROIs and ground-truth.

In summary, we propose an interaction-centric graph parsing network (iCGPN) in this paper. Firstly, for each detected human bounding box in images, the iCGPN constructs a fully-connected graph where a human box is treated as the center (the central node), and other boxes are the semantic nodes. During the training and inference stages, we enumerate each centric node with other surrounding nodes (semantics nodes) to construct a graph network. In this way, not only the dominant role of the human can be highlighted, but also richer HOI information related to the central node can be explored. Secondly, for semantic nodes, via introducing the attention mechanism, we propose an interaction-centric module to selectively extract semantic features to enrich node representations. Meanwhile, a multi-relation GCN is designed to measure the relationship between nodes and update node features with different edge information. Finally, a readout function is used to predict interactions according to node representations. Moreover, we adopt multi-IoU random shift for data augmentation to enable the proposed model to adapt to actual detection scenes.

To summarize, our contributions are as follows,

- By introducing a multi-relation GCN, an interaction-centric graph parsing network (iCGPN) is proposed to highlight the dominant role of humans in Human-Object Interactions.
- We propose a novel semantic node representation, in which interaction-related semantic features and meaningful object information are integrated.
- To enhance the generalization capability of iCGPN, a new data augmentation strategy based on multi-IoU random shift is designed during the training.
- The experimental results on the V-COCO [10] and HICO-DET [3] datasets show that the iCGPN achieves very competitive results in comparison with state-of-the-art graph-based methods, demonstrating the effectiveness of the proposed method.

II. Related Work

Object detection. Object detection [12, 15, 22, 34] plays an important role in HOI detection, where the human and object instances in images are firstly detected, and subsequently, the interactions between all pairs of people and objects are identified. [12, 22] are corresponding bounding box based two-stage detectors with accurate detection results. However, [15, 34] used keypoint estimation to find center points and regresses to object location so as to be faster than two-stage methods. In this paper, we use CenterNet [34] and the Hourglass-104 [15] models as the feature extractor and backbone network respectively.

Graph Neural Network (GNN). GNN has attracted more attention, and lots of graph models [4, 11, 14, 25] were proposed to capture the interdependent relations between nodes, in which node features could be dynamically updated by continuously obtaining information flow from neighboring nodes. Kipf, et al. [14] introduced the convolution operation into the graph network, and became the basis of many complex GNNs. Hamilton, et al. [11] generated the embedding vector of the target node by learning a function that aggregated neighboring nodes. Chen, et al. [4] propose a novel hierarchical graph neural network for few-shot learning to explore multi-level relationships. Cucurull, et al. [25] introduced the attention

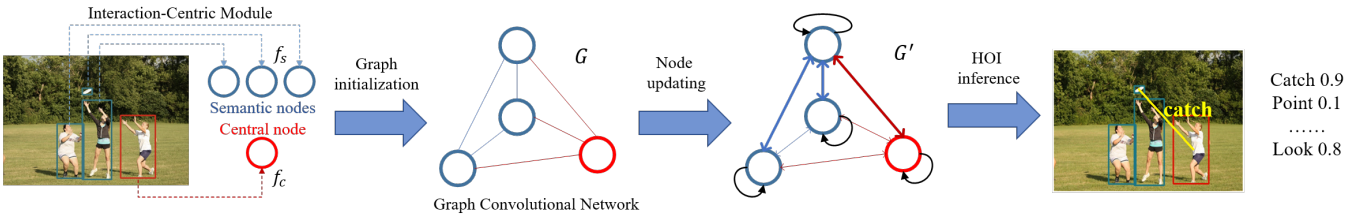

Fig. 3. The overview of iCGPN. After initializing the graph by using the central node (red node) and semantic nodes (blue nodes), a GCN is applied to update node features according to the connectivity matrix and predict the final HOI labels.

mechanism into the GNN, thus the weight aggregating feature of neighboring nodes could be determined. In this paper, considering semantic nodes representation has used visual features extracted by the attention mechanism, our proposed interaction-centric graph parsing network is elaborately constructed based on graph convolutional networks (GCNs) [14], not graph attention networks (GATs) [25].

HOI detection. HOI aims to deal with human-centric visual relationship detection. [2, 9] proposed a multi-stream model combining appearance features of detected human and object, and spatial layout features between human and objects. In [26], human pose information was also utilized to extract detailed local appearance cues for HOI detection. And some recent works [19, 20] paied more attention to extracting semantic information. Moreover, the visual attention mechanism was introduced to learn human and object features related to image contents. Gao, et al. [8] proposed an instance-centric attention network to refine the pairwise features. Wang, et al. [28] introduced context-aware human and object appearance features to learn important contextual features. Similar to them, we introduce an attention mechanism like a transformer to extract visual context features. In addition, the GNN was also applied to model the interactive relationship in HOI detection. Qi, et al. [21] constructed a homogeneous graph, in which humans and objects are equally treated, i.e., nodes and edges in graphs are looked as the same type, thus the leading role of humans was ignored. Zhou, et al. [33] proposed the object-bodypart graph and human-bodypart graph to assemble body part contexts for HOI detection. Lin, et al. [18] performed relation reasoning on human-object pairs with a graph network, where each graph node contained the information of each human-object pair, thus the relationships of objects were ignored. Zheng, et al. [27] designed a heterogeneous graph that models humans and objects as different kinds of nodes, yet the subjectivity of humans is still ambiguous because the conductions of different human nodes representation and the inter-class relations between human-object nodes are the same. Gao, et al. [7] constructed a human-centric and an object-centric HOI subgraph instead of full-connected graph and applied GAT [25] to aggregate the contextual information, but it also ignored the contextual cues among objects.

Existing graph-based methods have achieved significant results in HOI detection, however, as mentioned before, they have three main issues: 1) the leading role of humans is not obvious; 2) the relationships among objects are neglected; 3) The node representations rely on appearance features is weak for HOI inference. As a result, in this paper, an interaction-centric graph parsing network (iCGPN) is constructed to learn related HOIs.

III. Methodology

Fig.3 shows the framework of proposed iCGPN, which mainly contains three parts: graph initialization, node updating and HOI inference. Aiming to detect the interaction of human in the red box, we conduct a fully-connected graph with the central human node and three semantic nodes (two humans and one frisbee). In the initialization, for the central node, we construct a fully-connected graph G. Specifically, an interaction-centric module is applied to generate semantic node features f_s. Considering that human appearance contains information of the posture and action, we simply use a residual block followed by a Global Average Pooling (GAP) layer to extract central node representations f_c. Then, a link function is used to measure the connectivity between different graph nodes, and subsequently, according to predicted relationships, a GCN is applied to the fully-connected graph to update node features. Finally, based on the updated graph G', HOI labels are predicted using node features.

A. *Interaction-Centric Module*

To combine semantic interaction features with useful object information for semantic nodes, as shown in Fig. 4, an interaction-centric module is proposed, and it has two key components. Given the detected object instances by CenterNet [34], we apply global information sub-module to obtain interaction-related visual features. And meanwhile the node information sub-module is used to aggregate spatial and category information for HOI detection.

CenterNet proposed by Zhou et al. [34] was used for object detection here with the Hourglass-104 [15] backbone to generate bounding boxes for all human and object instances in an image. CenterNet models an object as a single point — the center point of its bounding box and regresses to its location. Hourglass-104 backbone is adopted to extract image features.

*1) **Global information sub-module**:* Considering that besides appearance features, the semantic node representations of objects should pay attention to semantic features, containing interactions between central and semantic nodes. Therefore, inspired by the transformer [24], a global information sub-module, in which central node boxes and semantic node boxes are regarded as the instance inputs respectively, is proposed to find the interaction-related features in the entire image.

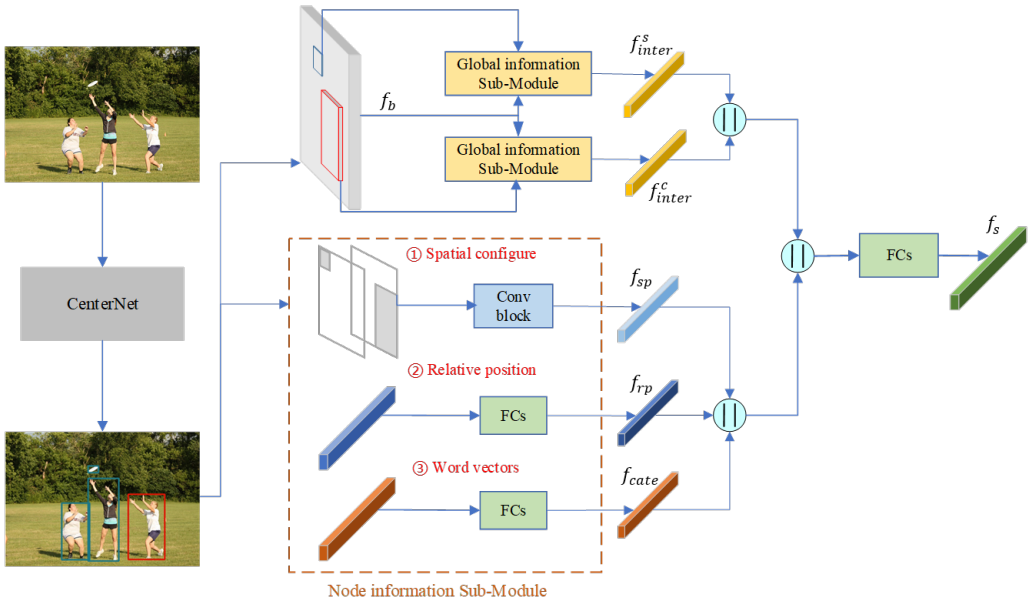

Fig. 4. Illustration of the interaction-centric module. || means concatenation. The backbone feature f_b is extracted from Hourglass-104 in CenterNet. f^c_{inter} and f^s_{inter} denote the features obtained from the central node and semantic nodes respectively. Two global information sub-modules share the parameters. The proposed node information sub-module (marked by dotted box) includes three parts: spatial configure, relative position, and word vectors, which are used to integrate spatial layout features and object categories. By concatenating all features, the interaction-centric module generates semantic node features f_s.

Fig.5 provides an overview of global information sub-module. Here, positional encoding and multi-head attention [24] are introduced. Firstly, we design two different positional encodings for the node feature and backbone feature. The conduction of node positional encoding is followed. After resizing to a fixed size, the original image region is taken as a reference, then a binary spatial mask is created by filling the bounding box of the node with 1. Afterwards, one-channel binary images are fed into a convolutional block to get the spatial feature f^{sp}_n. In addition, we use a 4-dimensional position vector $r_{n|b}$ as the input of the fully-connected layer to get a position feature f^{pos}_n of the node. The position vector can be written as:

$$r_{n|b} = \{\frac{x_n - x_b}{w_b}, \frac{y_n - y_b}{h_b}, log\frac{w_n}{w_b}, log\frac{h_n}{h_b}\} \quad (1)$$

where $(x_n,\ y_n)$ and $(x_b,\ y_b)$ are central coordinates of the node bounding boxes and original images, and $(w_n,\ h_n)$ and $(w_b,\ h_b)$ are their widths and heights. f^{pos}_n will generate more distinctive location features of the node, as f^{sp}_n maybe inaccurate due to spatial mask encoding errors caused by resizing operation. Thus, we add f^{pos}_n to f^{sp}_n and input it to a convolutional block to get the final node positional encoding f^n_{pe}, which supplies location information to node appearance features. It will be beneficial in HOI detection. For example, in a plays baseball scene, the appearance features of a baseball near a baseball bat should be different from a baseball falling on the ground, thereby position information provides a strong clue for accurate HOI detection.

Inspired by the successful applications [23, 29] of relative positional encodings in vision tasks with Transformer-based models, we simply add two learnable parameters $R^{H\times 1}_h$ and $R^{1\times W}_w$, whose shape respectively equals to the height and width of the backbone feature map after RoI align [12], to get backbone positional encoding with shape of $H \times W$. And then we obtain visual feature with positional awareness with the fusion of backbone feature map and positional encoding. Differently, we apply the supervision from the loss of HOI to guarantee that semantic node representations are interaction-related. Thus, the learnable backbone positional encoding also introduces spatial clues to capture interaction-related features. The visualization of positional encoding shows that it is effective to highlight image regions that are likely to contain human-object interaction instances.

To obtain multiple contextual appearance features, the multi-head attention is introduced to the global information sub-module. Specifically, as shown in Fig.5, we firstly obtain query features, key features and value features. Then, these features are divided into S slices according to the channels and mark them as $\{q_i\}^S_{i=1}$, $\{k_i\}^S_{i=1}$, $\{v_i\}^S_{i=1}$ respectively. Finally, we apply scaled dot-product operation for each pair of features separately and concatenate results from each $q-k-v$ pair to obtain the final output feature f_{inter} (f^c_{inter} for central nodes and f^s_{inter} for semantic nodes).

$$f_{inter} = FC[||^S_{i=1} softmax(q^T_i k_i)v_i] \quad (2)$$

where $FC[.]$ denotes fully-connected layers, and || indicates concatenation. [24] indicates that multi-head attention allows the model to jointly attend to information from different representation subspaces. And the feature maps of different channels obtained from CNN with many filters correspond to different feature subspaces. Thus, we split features based on channels and compute each $q-k-v$ pair on each feature map slice. In experiment, the effectiveness of aggregating features related to HOI from multiple representation subspaces with multi-head attention will be demonstated.

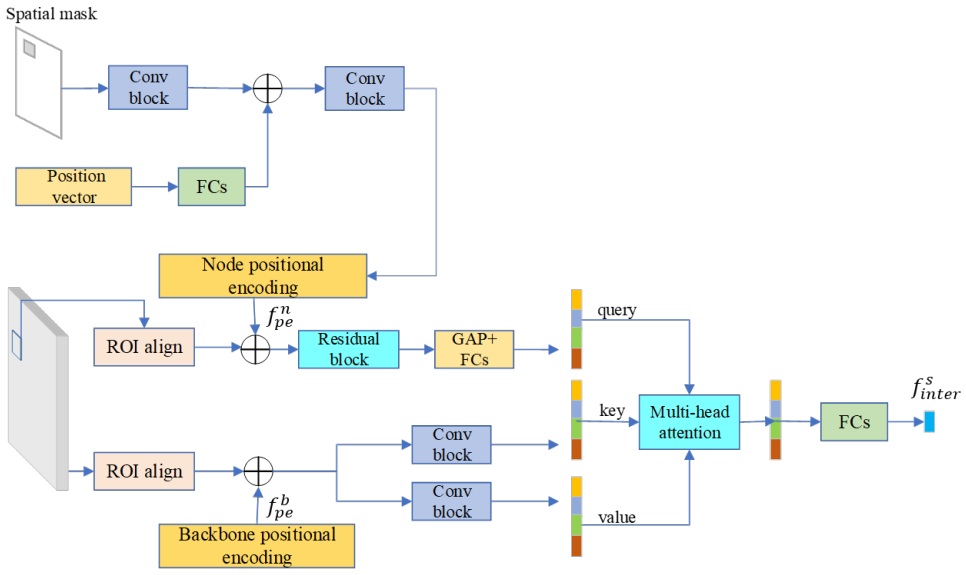

Fig. 5. The overall architecture of global information sub-module. $\bigoplus$ means add operation. The multi-head attention is introduced to obtain multiple contextual appearance features. Two different positional encodings are designed to represent node features and backbone features respectively. The node positional encoding module, where spatial features from spatial masks and location features from position vectors are combined to get more accurate and distinctive positional encoding.

*2) **Node information sub-module**:* As shown in Fig. 4, three different features related to HOI are combined in this sub-module. Previous works [8, 9, 16, 26] indicated that the spatial relationships of human-object pairs could provide strong cues for interaction detection. In our implementation, we use a two-channel binary union region mask to encode spatial relationships between central nodes and semantic nodes. Then, a convolutional network is applied to extract spatial layout features f_{sp} from spatial relation maps. In addition to spatial features, two important object-related factors contribute to HOI detection, namely, relative positions and object categories. Inspired by InteractNet [9], a 4D relative position encoding $r_{s|c}$, as shown in Equation (3), is obtained as the input of fully-connected layers to extract the relative position features f_{rp}.

$$r_{s|c} = \{\frac{x_s - x_c}{w_c}, \frac{y_s - y_c}{h_c}, log\frac{w_s}{w_c}, log\frac{h_s}{h_c}\} \tag{3}$$

where (x_s, y_s) and (x_c, y_c) are central coordinates of the bounding boxes of semantic node and central node respectively, and (w_s,h_s) and (w_c, h_c) are widths and heights of boxes. For category features f_{cate}, we use Bert [6] to extract word vectors of object categories and then the 768 dimensional word vectors are fed into a multi-layer perceptron (MLP) with a single 96 dimensional hidden layer with ReLU to extract category representations.

Through two sub-modules, different extracted features (f^c_{inter}, f^s_{inter}, f_{sp}, f_{rp} and f_{cate}) are concatenated, and then a MLP with one hidden layer with 512 neurons is used to obtain the final semantic node representations f_s. The new semantic node representations not only contain multiple contextual features complementary to the appearance features, but also add useful information, such as spatial configure, relative positions and category information for HOI detection, which is insensitive to appearance variations.

B. Graph Convolutional Network

After extracting features for graph nodes, a GCN is built to update node representations and predict interaction labels.

Similar to GPNN [21], a link function $L(\bullet)$ is constructed to measure connectivity between nodes. $L(\bullet)$ considers the central node features f_c and semantic node features f_s as the inputs, and outputs an adjacency matrix $A \in [0,1]^{|V|\times|V|}$ as follows:

$$A_{ij} = sigmoid(L([f_i, f_j])) \tag{4}$$

where A_{ij} denotes the (i, j)-th entry of the adjacency matrix A, and f_i and f_j are the i-th and j-th node features respectively. [.., ..] indicates concatenation, $L(\bullet)$ is implemented by a simple convolutional block.

Considering different roles that central nodes and semantic ones play in human-object interaction, we divide the adjacency matrix into four parts: 1) identity matrix $A_0 = I$ is used for self-update; 2) semantic-semantic relation matrix A_1 is used to update semantic nodes with other semantic node features. The nodes can gather more messages from other relevant nodes by learning with context among semantic nodes, and all semantic nodes construct the semantic information of the entire scene; 3) semantic-central relation matrix A_2 is applied for updating semantic nodes with central node features. Contextual features of semantic nodes can be supplemented by central node appearance features, which contain some useful clues for interaction detection, such as posture; 4) central-semantic relation matrix A_3 is adopted to update central nodes with semantic node features. The context from surrounding semantic nodes provides abundant scene information for the central node to

better understand and determine the HOIs. In addition, A_2 and A_3 indicate the relationships between different node types. Thus, they are also expected to determine whether the nodes have interactions to strengthen the relations between nodes, which are interactive. According to this relation partitioning strategy, node features are updated with the following formula:

$$f_{post} = \sum_j \Lambda_j{}^{-\frac{1}{2}} A_j \Lambda_j{}^{-\frac{1}{2}} f_{pre} W_j \tag{5}$$

where Λ is a diagonal matrix used for graph normalization, and $\Lambda_j^{ii} = \sum_k (A_j^{ik}) + \alpha$. Here, we set $\alpha = 10^{-5}$ to avoid empty columns. f_{pre} and f_{post} are node features before and after update, and W_j is weight matrix for different feature transformations.

Lastly, the node features are fed into fully-connected layers to output interaction labels. Here, central node features are used to predict all related actions of the central node, whereas semantic node features are applied to infer central-semantic node pair related actions, showing the uniqueness of human nodes in graphs. The final action score of the object node is calculated by the following:

$$s_{c,s}^a = \left(s_c^a + s_s^a * r_{cs}\right)/2 \tag{6}$$

where s_c^a and s_s^a are the output score of the central node and semantic node for the action a, respectively; r_{cs} is the relation score calculated from the relation matrix. Then, the action-related object is confirmed by the following:

$$b_{s^*} = argmax_{b_s} s_s^a * r_{cs} \tag{7}$$

C. *Multi-IoU random shift*

Because the object detection results during inference may not be as excellent as those during training, thus the unbalance data distribution is serious. To deal with this problem and to improve the generalization ability of the network, a multi-IoU random shift is proposed to generate augmented bounding boxes to satisfy approximately uniform distribution of IoU with GT boxes between 0.5 and 1.0, as shown in Fig.6(b). In our implementation, two boxes (blue boxes in Fig.6(a)) with the same center and the same aspect ratio as the GT box (red box in Fig.6(a)) are generated, and the IoU of these boxes with the GT box is p. Augmented bounding boxes (green dotted box in Fig.6(a)) should satisfy the requirement of including the boundary of the smaller box but not exceeding the one of the larger box. We find that the IoU of all generated boxes satisfy approximately normal distribution of IoU with GT boxes between p and 1. Thus, a large number of generated boxes satisfy the IoU $(p+1)/2$ with the GT box. Therefore, to obtain a large number of augmented bounding boxes with IoU as q, the two generated boxes with the same centre and aspect ratio as the GT box should satisfy the IoU $2q+1$ with the GT box. During training, q will sample from a uniform distribution between 0.5 and 1.

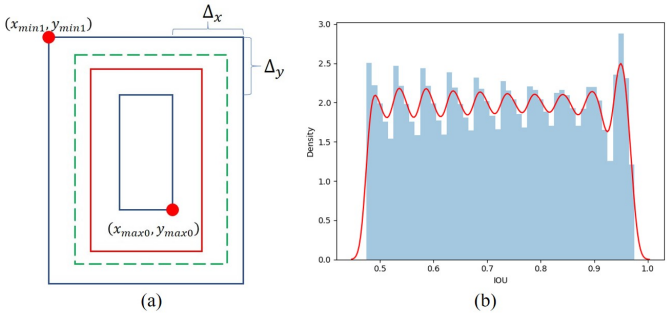

Fig. 6. (a) Overview of multi-IoU random shift. The red box is GT and two blue ones have the same center and aspect ratio with the GT. The green dotted box is an augmented box. (b) An approximate uniform IoU distribution after data augmentation.

$$\begin{aligned}
x_{max0} &= x + 0.5 * w * \sqrt{p} \\
y_{max0} &= y + 0.5 * h * \sqrt{p} \\
x_{min1} &= max(0,\ x - 0.5 * w * \frac{1}{\sqrt{p}}) \\
y_{min1} &= max(0,\ y - 0.5 * h * \frac{1}{\sqrt{p}}) \\
\Delta_x &= 0.5 * w * \frac{1}{\sqrt{p}} - 0.5 * w * \sqrt{p} \\
\Delta_y &= 0.5 * h * \frac{1}{\sqrt{p}} - 0.5 * h * \sqrt{p}
\end{aligned} \tag{8}$$

Fig.6(a) shows an overview of our method. The masks in Fig.6(a) can be calculated according to Equation (8), where (x, y) is the central coordinate of GT boxes, (w, h) are widths and heights of GT boxes, respectively. According to Equation (9), where $U(0,1)$ denotes the uniform distribution between 0 and 1, augmented bounding boxes can be obtained, and the green dotted box in Fig.6(a) shown an instance, whose coordinates are (x_{min}, y_{min}, x_{max} and y_{max}).

$$\begin{aligned}
x_{min} &= \Delta_x * U(0,1) + x_{min1} \\
y_{min} &= \Delta_y * U(0,1) + y_{min1} \\
x_{max} &= \Delta_x * U(0,1) + x_{max0} \\
y_{max} &= \Delta_y * U(0,1) + y_{max0}
\end{aligned} \tag{9}$$

D. *The Loss*

Considering HOI detection is a multilabel task, a multilabel loss for interaction classification is used. Here, an asymmetric loss [1] L_{HOI} for final interaction classification is applied to solve the imbalance problem of positivenegative samples in HOI datasets. Meanwhile, to guarantee that semantic node representations focus on interaction-related semantic features, we use semantic node features to predict human-object interactions and apply another asymmetric loss L_{inter} as an intermediate loss. In relation to matrix learning, the L1 loss L_r for human-object relation training is adopted. In summary, the objective function L is the weighted sum of the interaction

classification loss L_{HOI}, L_{inter} and interactiveness loss L_r, as follows:

$$L = L_{HOI} + \alpha * L_{inter} + \beta * L_r \tag{10}$$

where α and β are the hyper-parameters and are set to 0.1 in the experiments.

IV. Experiments

To evaluate the effectiveness of the proposed method, we conduct extensive experiments on the HICO-DET and V-COCO datasets.

A. Datasets

Two HOI datasets, i.e., HICO-DET [3] and V-COCO [10] are used to evaluate the performance. HICO-DET includes 47,776 images (38,118 images for training and 9,658 ones for testing), which are labeled into 80 object categories, 117 verb classes and 600 HOI categories. HICO-DET also provides more than 150K annotated instances of human-object pairs. V-COCO dataset contains 10,346 images (2,533 images for training, 2,867 images for validation, and 4,946 ones for testing) and 16,199 human instances, which are annotated with 80 object categories and 29 verb classes.

B. Metrics

In our experiments, the role means average precision (role mAP [10]) is used as the evaluation metric. A $\langle human, action, object \rangle$ triplet is considered as a true positive if the predicted action matches the ground-truth, and both predicted human and object bounding boxes have $min(IOU_h, IOU_o) \geq 0.5$ with reference to GT boxes. For the cases when there is no object (human only) in VCOCO, a prediction is correct if the corresponding bounding box for the object is empty in Scenario 1 and the bounding box of the object is not considered in Scenario 2.

TABLE I
PERFORMANCE COMPARISON ON THE V-COCO TEST SET. THE BEST AND THE RUNNER-UP ARE LABELED IN **BOLD** AND UNDERSCORED RESPECTIVELY.

Method	role mAP (Sc 1)	role mAP (Sc 2)
GPNN[21](ECCV2018)	44.0	-
Xu *et al.*[30](CVPR2019)	45.9	-
Li *et al.*[16](CVPR2019)	47.8	-
RPNN[33](ICCV2019)	36.7	47.5
PMFNet[26](ICCV2019)	52.0	-
AGRR[18](IJCAI2020)	48.1	-
Contextual HGNN[27](ECCV2020)	52.7	-
In-GraphNet[31](IJCAI-PRICAI2020)	48.9	-
SIGN[32](ICME2020)	53.1	-
DRG[7](ECCV2020)	51.0	-
iCGPN(**ours**)	**53.8**	**62.7**

C. Implementation Details

We use CenterNet [34] with a backbone of Hourglass-104 [15] to generate human and object bounding boxes. In test, human and object boxes with scores higher than 0.3 will be kept. We initialize our appearance feature backbone network with the COCO pretrained weight from CenterNet. During training, for each central node, semantic nodes are randomly sampled, i.e., includes 5 positive samples and up to 15 negative samples. Because the number of detected node bounding boxes is different in different images, thereby this strategy can improve the performance and robustness of our model. Note that random sampling is not performed in testing. For HICO-DET, we train the model for 15 epochs using SGD with a learning rate of 1e-4, a momentum of 0.9 and a weight decay of 1e-5. For V-COCO, the model is trained for 30 epochs with a learning rate of 1e-5. All experiments are conducted on a single Nvidia GeForce 1080Ti GPU with PyTorch. Training our network on V-COCO requires 0.5 hours for each epoch, and HICO-DET requires 3.5 hours. In testing, our model runs at 4 fps for V-COCO and HICO-DET.

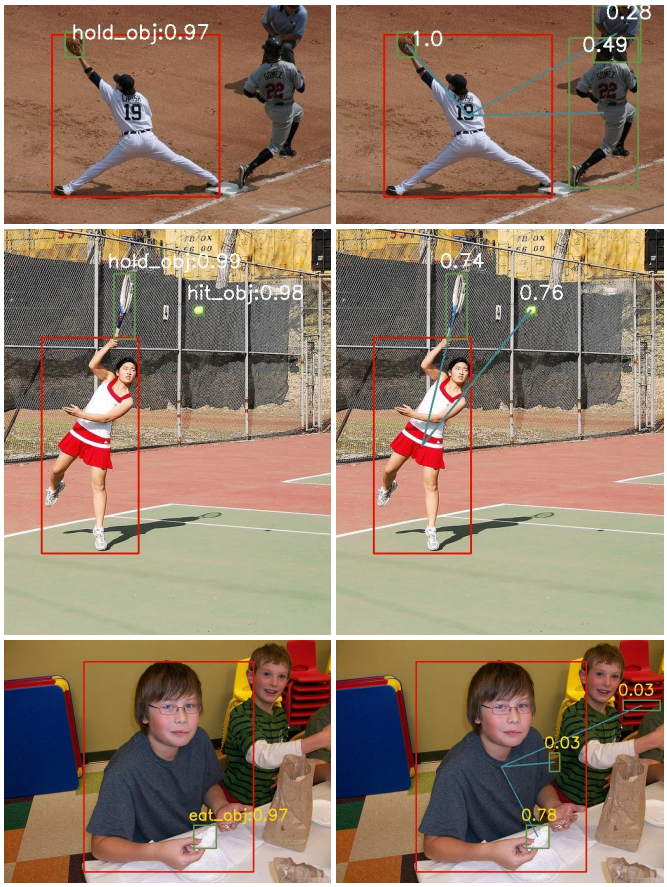

Fig. 7. Several visualization results of HOI detection. Blue lines connect all human-object pairs, and relation scores give the confidences.

D. Quantitative evaluation

The overall quantitative results of role mAP on V-COCO and HICO-DET are listed in Table I and Table II respectively. For the V-COCO dataset, our proposed iCGPN achieves the best performance among state-of-the-art graph-based methods. In general, many methods introduce pose information as additional feature to more accurately classify the HOI, while it requires complex data preprocessing. For example, compared with PMFNet[26], our method without human pose information still achieves a significant performance gain (1.8%).

TABLE II
PERFORMANCE COMPARISON ON THE HICO-DET TEST SET.

	Default			Known Object		
Method	Full	Non Rare	Rare	Full	Non Rare	Rare
GPNN[21](ECCV2018)	13.11	14.23	9.34	-	-	-
Xu *et al.*[30](CVPR2019)	14.70	15.13	13.26	-	-	-
Li *et al.*[16](CVPR2019)	17.03	18.11	13.42	-	-	-
RPNN[33](ICCV2019)	17.35	18.71	12.78	-	-	-
PMFNet[26](ICCV2019)	17.46	18.00	15.65	20.34	21.20	17.47
AGRR[18](IJCAI2020)	16.63	18.22	11.30	19.22	20.61	14.56
Contextual HGNN[27](ECCV2020)	17.57	17.78	16.85	21.00	21.08	20.74
In-GraphNet[31](IJCAI-PRICAI2020)	17.72	19.31	12.93	-	-	-
SIGN[32](ICME2020)	17.51	18.53	15.31	20.49	21.51	17.53
DRG[7](ECCV2020)	19.26	19.71	**17.74**	**23.40**	**23.89**	**21.75**
iCGPN(**ours**)	**19.40**	**20.19**	16.76	21.81	22.63	19.08

Furthermore, the results show that our iCGPN outperforms all graph-based methods such as GPNN [21], RPNN [33], AGRR [18], In-GraphNet [31], SIGN [32] and DRG [7]. Particularly, we improve dramatically by 0.7% to SIGN [32], which gained the second best performance. These results demonstrate the effectiveness of our approach.

For the HICO-DET dataset, the results show that our method achieves the best performance in full setting and non-rare setting of Default mode, as well as achieves very competitive results on Known Object mode. The results of rare setting is slightly low because the long tail distribution is very serious in HICO-DET dataset. Particularly, iCGPN outperforms most graph-based methods by a significant margin, such as GPNN [21], RPNN [33], AGRR [18], In-GraphNet [31] and SIGN [32].

E. Subjective evaluation

Visualizing the HOI detection results of iCGPN. In Fig.7, the detected persons and objects are drawn by red and green bounding boxes, and the predicted action labels and scores in the left images are also annotated, and meanwhile, the relation scores, indicating the interaction between each human-object pair, are listed in the right. The first row in Fig.7 shows a human holding a baseball glove; thus, the relation score between the human and baseball gloves is evidently higher than other object nodes without interactions. In the middle row, the woman is holding a tennis racket to hit a sports ball. Thus, the relation scores between the human node and two object nodes are similar and high. The third row shows a boy eating a cake, its relation score with the cake is far higher than two objects with low object detection confidence, and interestingly, although other two nodes are detected, they are inaccurate, thus the confidences are very low. The above results show that the relation module can effectively exclude low-confidence semantic nodes.

Analysis of positional encoding in global information sub-module. Fig.8(a) visualizes typical node positional encodings. We find that obvious differences appear in the positional encoding of detected nodes if positions are different, and the high bright values are the responses of location encoding of nodes, which can effectively add location information to the appearance features of the nodes. As shown in Fig.8(b), the highlighted region really pops up a human-object interaction, and will enhance the effectiveness of feature representation of nodes.

F. Ablation Analysis

TABLE III
ABLATION STUDY OF THE CORE COMPONENTS IN GLOBAL INFORMATION SUB-MODULE.

Method	role mAP (Sc 1)
w/o Global information Sub-Module	49.7
w/o Node information Sub-Module	46.9
w/o Multi-relation GCN	52.0
w/o Multi-IoU random shift	53.1
iCGPN	**53.8**

Effectiveness of all modules in iCGPN. In Table III, we evaluate the contributions of different modules. Evidently, the best performance is obtained via combining all modules. The comparisons on the role mAP of different verbs with or without data argumentation situations imply the effectiveness of the data argumentation, as shown in Fig.9. Specifically, as listed in rows 1 and 5 of Table III, the proposed design of the global information sub-module improves role mAP by 4.1%, indicating that the learned interaction features are more discriminative in HOI detection. Similarly, because the node information sub-module can learn the information of spatial layout, object relative location and category to coordinate with interaction-centric features, the sub-module can improve the overall performance from 46.9% to 53.8%. In addition, we conduct four different types of feature transformation, corresponding to four edge relationships in GCN according to the roles of nodes. To verify the effectiveness of our proposed method, we conduct experiments on the graph model that updates node features with the same feature transformation. Table III shows that with the multi-relation GCN, the mAP increases 1.8%, indicating that diversified contextual information is very valuable in improving the feature representation for HOI. These experiments demonstrate the effectiveness of our proposed method.

Effectiveness of core components in global information sub-module. Table IV shows that the mAP approximately decreases by 1% in V-COCO when the multi-head attention or positional encoding is not adopted. For examining the influence of multi-head attention, we do not split features

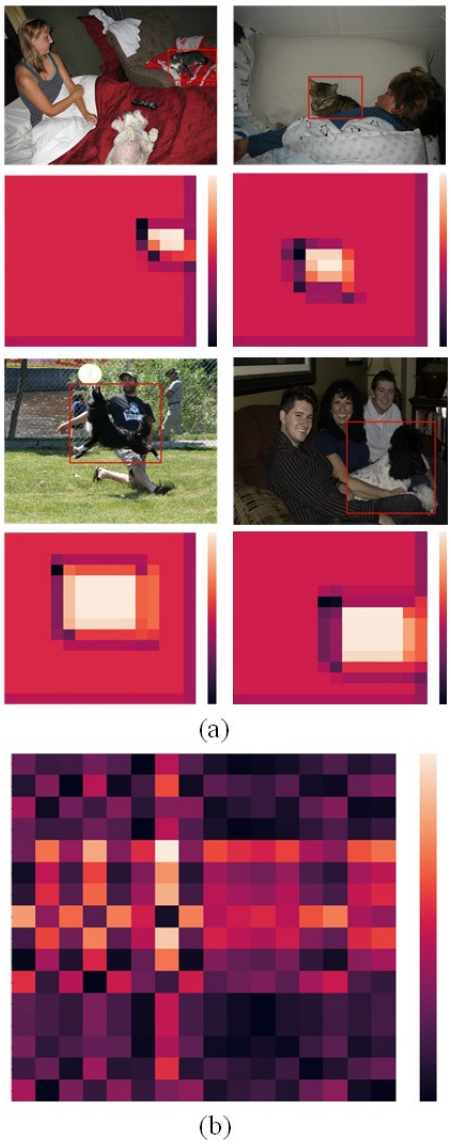

Fig. 8. (a) Node positional encoding visualization. The detected object is annotated in red, and the encoding heatmaps below show the location responses of the object. (b) Backbone positional encoding visualization. The heatmap has a high response value in the region where the interactions often occur in datasets.

TABLE IV
ABLATION STUDY OF THE CORE COMPONENTS IN GLOBAL INFORMATION SUB-MODULE.

Method	role mAP (Sc 1)
w/o multi-head attention	52.6
w/o positional encoding	52.9
iCGPN	**53.8**

into many slices according to channels (denoted as w/o multi-head attention). The experimental result indicates that nodes can selectively learn multiple contextual features with strong relevance to HOI based on multi-head attention, and avoid gather useless messages from entire images. In addition, location information, provides a clue for HOI, is supplied into appearance features with positional encoding, thereby improves model performance. Experimental results show that

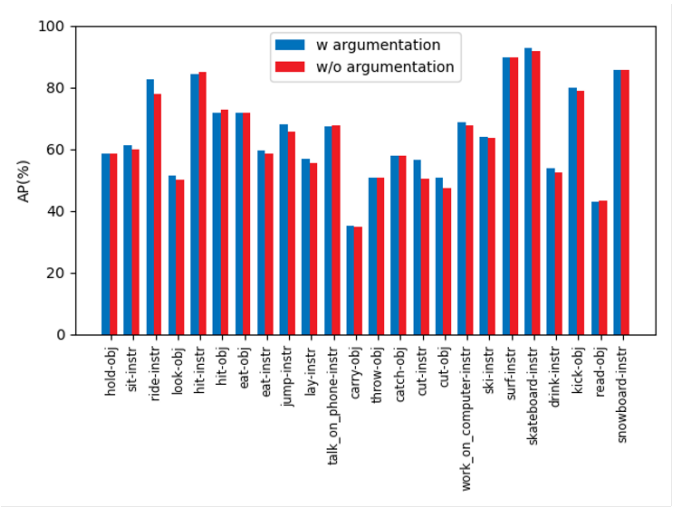

Fig. 9. Comparisons of the role mAP (Sc 2) of different verbs under two situations: train network with or without data argumentation.

the global information sub-module can effectively capture the interactive information between the central node and semantic nodes.

V. CONCLUSION

In this paper, a novel interaction-centric graph parsing network (iCGPN) is proposed for HOI detection. Through mining interaction-related features and preserving useful object information including relative position, spatial layout and object category, a new semantic node representation is applied. Since human-related semantic features and positional features are both considered in semantic nodes, the dominant role of humans in human-object interactions, named as central nodes, is highlighted. Furthermore, a hierarchical random shift training strategy is proposed to augment data of the training set, which enhances the generalization ability of the proposed model. The extensive experimental results demonstrate the effectiveness of the proposed method.

ACKNOWLEDGMENT

This work is supported by Chinese National Natural Science Foundation (62076033, U1931202), and The National Key Research and Development Program of China (2020YFB2104600).

REFERENCES

[1] Emanuel Ben-Baruch, Tal Ridnik, Nadav Zamir, Asaf Noy, Itamar Friedman, Matan Protter, and Lihi Zelnik-Manor. Asymmetric loss for multi-label classification. *arXiv preprint arXiv:2009.14119*, 2020.

[2] Yu-Wei Chao, Yunfan Liu, Xieyang Liu, Huayi Zeng, and Jia Deng. Learning to detect human-object interactions. In *2018 ieee winter conference on applications of computer vision (wacv)*, pages 381–389. IEEE, 2018.

[3] Yu-Wei Chao, Zhan Wang, Yugeng He, Jiaxuan Wang, and Jia Deng. Hico: A benchmark for recognizing human-object interactions in images. In *Proceedings of*

the IEEE International Conference on Computer Vision, pages 1017–1025, 2015.

[4] Cen Chen, Kenli Li, Wei Wei, Joey Tianyi Zhou, and Zeng Zeng. Hierarchical graph neural networks for few-shot learning. *IEEE Transactions on Circuits and Systems for Video Technology*, pages 1–1, 2021.

[5] Xiaoyu Chen, Hongliang Li, Qingbo Wu, King Ngi Ngan, and Linfeng Xu. High-quality r-cnn object detection using multi-path detection calibration network. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(2):715–727, 2021.

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[7] Chen Gao, Jiarui Xu, Yuliang Zou, and Jia-Bin Huang. Drg: Dual relation graph for human-object interaction detection. In *European Conference on Computer Vision*, pages 696–712. Springer, 2020.

[8] Chen Gao, Yuliang Zou, and Jia-Bin Huang. ican: Instance-centric attention network for human-object interaction detection. *arXiv preprint arXiv:1808.10437*, 2018.

[9] Georgia Gkioxari, Ross Girshick, Piotr Dollár, and Kaiming He. Detecting and recognizing human-object interactions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8359–8367, 2018.

[10] Saurabh Gupta and Jitendra Malik. Visual semantic role labeling. *arXiv preprint arXiv:1505.04474*, 2015.

[11] William L Hamilton, Rex Ying, and Jure Leskovec. Inductive representation learning on large graphs. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 1025–1035, 2017.

[12] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[13] Jian-Fang Hu, Wei-Shi Zheng, Jianhuang Lai, Shaogang Gong, and Tao Xiang. Exemplar-based recognition of humanobject interactions. *IEEE Transactions on Circuits and Systems for Video Technology*, 26(4):647–660, 2016.

[14] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

[15] Hei Law and Jia Deng. Cornernet: Detecting objects as paired keypoints. In *Proceedings of the European conference on computer vision (ECCV)*, pages 734–750, 2018.

[16] Yong-Lu Li, Siyuan Zhou, Xijie Huang, Liang Xu, Ze Ma, Hao-Shu Fang, Yanfeng Wang, and Cewu Lu. Transferable interactiveness knowledge for human-object interaction detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3585–3594, 2019.

[17] Yue Liao, Si Liu, Fei Wang, Yanjie Chen, Chen Qian, and Jiashi Feng. Ppdm: Parallel point detection and matching for real-time human-object interaction detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 482–490, 2020.

[18] Xue Lin, Qi Zou, and Xixia Xu. Action-guided attention mining and relation reasoning network for human-object interaction detection. In *IJCAI*, pages 1104–1110, 2020.

[19] Hanchao Liu, Tai-Jiang Mu, and Xiaolei Huang. Detecting humanobject interaction with multi-level pairwise feature network. *Computational Visual Media*, 7(2):229–239, 2021.

[20] Yang Liu, Qingchao Chen, and Andrew Zisserman. Amplifying key cues for human-object-interaction detection. In *European Conference on Computer Vision*, pages 248–265. Springer, 2020.

[21] Siyuan Qi, Wenguan Wang, Baoxiong Jia, Jianbing Shen, and Song-Chun Zhu. Learning human-object interactions by graph parsing neural networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 401–417, 2018.

[22] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28:91–99, 2015.

[23] Aravind Srinivas, Tsung-Yi Lin, Niki Parmar, Jonathon Shlens, Pieter Abbeel, and Ashish Vaswani. Bottleneck transformers for visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16519–16529, 2021.

[24] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[25] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. *arXiv preprint arXiv:1710.10903*, 2017.

[26] Bo Wan, Desen Zhou, Yongfei Liu, Rongjie Li, and Xuming He. Pose-aware multi-level feature network for human object interaction detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9469–9478, 2019.

[27] Hai Wang, Wei-shi Zheng, and Ling Yingbiao. Contextual heterogeneous graph network for human-object interaction detection. In *European Conference on Computer Vision*, pages 248–264. Springer, 2020.

[28] Tiancai Wang, Rao Muhammad Anwer, Muhammad Haris Khan, Fahad Shahbaz Khan, Yanwei Pang, Ling Shao, and Jorma Laaksonen. Deep contextual attention for human-object interaction detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5694–5702, 2019.

[29] Haiping Wu, Bin Xiao, Noel Codella, Mengchen Liu, Xiyang Dai, Lu Yuan, and Lei Zhang. Cvt: Introducing convolutions to vision transformers. *arXiv preprint arXiv:2103.15808*, 2021.

[30] Bingjie Xu, Yongkang Wong, Junnan Li, Qi Zhao, and Mohan S Kankanhalli. Learning to detect human-object

interactions with knowledge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

[31] Dongming Yang and Yuexian Zou. A graph-based interactive reasoning for human-object interaction detection. *arXiv preprint arXiv:2007.06925*, 2020.

[32] Sipeng Zheng, Shizhe Chen, and Qin Jin. Skeleton-based interactive graph network for human object interaction detection. In *2020 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2020.

[33] Penghao Zhou and Mingmin Chi. Relation parsing neural network for human-object interaction detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 843–851, 2019.

[34] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. Objects as points. *arXiv preprint arXiv:1904.07850*, 2019.

Wenhao Yang is currently a master candidate at the Beijing Key Laboratory of Network System and Network Culture, Beijing University of Posts and Telecommunications, Beijing, China. His research interests include computer vision and deep learning.

Guanyu Chen received the M.S. degree in information and communication engineering from Beijing University of Posts and Communications, Beijing, China, in 2021. In July 2021, He joined Beijing Academy of Blockchain and Edge Computing, as a researcher. His research interests focus on human action recognition, remote sensing data fusion and deep learning in computer vision.

Zhicheng Zhao is an associate professor of Beijing University of Posts and Telecommunications. He was a visiting scholar at School of Computer Science, Carnegie Mellon University from 2015 to 2016. His research interests are computer vision, image and video semantic understanding and retrieval. He has authored and coauthored more than 70 journal and conference papers.

Fei Su is a female professor of Beijing university of posts and telecommunications. She received the Ph.D. degree majoring in Communication and Electrical Systems from BUPT in 2000. She was a visiting scholar at electrical computer engineering department, Carnegie Mellon University from 2008 to 2009, Her current interests include pattern recognition, image and video processing and biometrics. She has authored and co-authored more than 70 journal and conference papers and some textbooks.

Hongying Meng (M10-SM17) received his Ph.D. degree in Communication and Electronic Systems from Xian Jiaotong University, Xian China. He is currently is a Reader at the Department of Electronic and Electrical Engineering, College of Engineering, Design and Physical Sciences, Brunel University London. Dr Meng has a wide research interests including digital signal processing, machine learning, human computer interaction, computer vision, image processing and embedded systems. He is a Senior Member of IEEE and an associate editor for IEEE Transactions on Circuits and Systems for Video Technology (TCSVT) and IEEE Transactions on Cognitive and Developmental Systems (IEEE TCDS) and Journal of Real-Time Image Processing (Springer).