Relation Parsing Neural Network for Human-Object Interaction Detection

Penghao Zhou and Mingmin Chi
Shanghai Key Laboratory of Data Science, Fudan Univeristy, Shanghai, China
{17210240267,mmchi}@fudan.edu.cn

Abstract

Human-Object Interaction Detection devotes to infer a triplet $\langle human, verb, object \rangle$ between human and objects. In this paper, we propose a novel model, i.e., Relation Parsing Neural Network (RPNN), to detect human-object interactions. Specifically, the network is represented by two graphs, i.e., Object-Bodypart Graph and Human-Bodypart Graph. Here, the Object-Bodypart Graph dynamically captures the relationship between body parts and the surrounding objects. The Human-Bodypart Graph infers the relationship between human and body parts, and assembles body part contexts to predict actions. These two graphs are associated through an action passing mechanism. The proposed RPNN model is able to implicitly parse a pairwise relation in two graphs without supervised labels. Experiments conducted on V-COCO and HICO-DET datasets confirm the effectiveness of the proposed RPNN network which significantly outperforms state-of-the-art methods.

1. Introduction

Deep convolutional neural networks [1, 2, 3] have led a rapid improvement for visual recognition tasks, including object detection [4, 5, 6] and human keypoint estimation [7, 8, 9]. Individual object detection is the basis of understanding images. To comprehend the content of the images, visual relationship recognition between individual instances is important and challenging. Human-Object Interaction (HOI) Detection [10, 11, 12, 13] is the related task which aims to detect human and objects, and meanwhile to reason the complex interactions between them. We present an HOI detection instance in Fig. 1(a), where a triplet $\langle human, verb, object \rangle$ is respectively detected by our method.

Most existing methods [14, 15, 16, 17] obtain great detection results by considering the pair of human feature and object feature as shown in Fig. 1(b), and combining the spatial relationship to detect HOIs. However, directly using the whole human feature with uniform attention might lose a detailed information. Since human interact with objects only through body parts, the region of the body parts are of the most important for HOI detection. Therefore, the contents of body parts should be further considered for

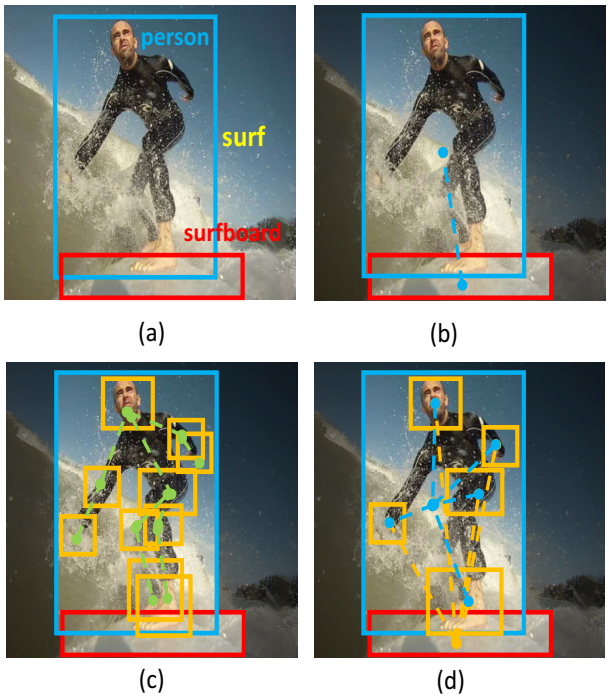

Figure 1. **Human-object interaction detection.** **(a)** Given an image, $\langle human \rangle$, $\langle verb \rangle$, and $\langle object \rangle$ are respectively detected to show that the person $(human)$ surfs $(verb)$ on the surfboard $(object)$, where the person box is marked in blue, and the object box is marked in red. **(b)** Recent methods [14, 15, 16, 17] directly use the pairs of the human and the object features to detect human-object interactions, where the blue line connects the human box and the object box to denote a pair of human and object. **(c)** the correlation between different body parts is exploited for HOI recognition in [18], where the body parts are marked in yellow, and the green lines denote the body-part pairs. **(d)** In contrast, our method exploits relations between *object-bodypart* (denoted with yellow lines) as well as *human-bodypart* (blue lines) for HOI detection. In this example, the strong relation between leg (body part) and surfboard (object) indicates that the person surfs on the surfboard.

modeling HOIs. To solve this problem, Fang et. al. [18] propose a pairwise body-part attention model to exploit the correlations between body parts as shown in Fig. 1(c). Different from HOI detection, HOI recognition is to evaluate the probabilities of certain interactions in a predefined HOI list. As mentioned in [18], the studies in cognitive science [19, 20] show that the visual attention of human is non-uniform and we tend to pay attentions to different body parts according to different contexts. Namely, we should pay different attention to different body parts for different objects and different human appearances.

To achieve this, we focus on the relationships between objects and body parts, as well as the human and body parts as shown in Fig. 1(d). For clarity, we abbreviate object and body part pair as *object-bodypart*, and the pair of human and body part as *human-bodypart* throughout the paper.

As illustrated in Fig. 1(d), the pair of surfboard and leg is an obvious cue to detect the interaction between the person and the surfboard. Therefore, inspired by [21], we construct an Object-Bodypart Graph for each person to capture the important relation implicit in the object-bodypart pair. Accordingly, the features of both object and body part can be refined through discovering certain object-bodypart (i.e., surfboard-leg in Fig. 1(d)) relations. Moreover, we tend to pay more attention to the leg part rather than overall human appearance. The attention-based Human-Bodypart Graph is built to parse the important relations of human-bodypart to refine the feature of human in the paper. In terms of the human-bodypart (i.e., human-leg) relation, the refined body part feature (i.e., leg) in the Object-Bodypart Graph can refine the human feature to tell what action is the human participating (surfing) and where the action is happening (near his leg).

Following the design in [16], we predict action and density estimation of target object location (where the action is happening) based on the refined human features, as the refined human features contain more clues to detect actions as well as estimate the density over the target object location for each action. Moreover, the interaction probabilities between the object and human for each action are predicted based on the refined object features. Along with the standard object/keypoints detection task, the RPNN is trained as a multi-task learning system (e.g. action classification, target object location regression and object classification/regression). By minimizing the end-to-end loss, our model can implicitly parse the relations of human-bodypart and object-bodypart.

In the paper, a novel model, i.e., Relation Parsing Neural Network (RPNN), is proposed to detect human-object interactions (HOIs). In particular, we address the HOI Detection problem by modeling the RPNN by two attention based graphs, i.e., Human-Bodypart Graph and Object-Bodypart Graph. Here, the Object-Bodypart Graph dynamically captures the relationship between body parts and the surrounding objects. The Human-Bodypart Graph infers the relationship between human and body parts, and assembles body part contexts to predict actions. These two graphs are associated through an action passing mechanism. The proposed RPNN is evaluated on two public datasets (i.e., V-COCO [13] and HICO-DET [15]). The experimental results confirm the effectiveness of the proposed RPNN network by achieving the state-of-the-art performance.

The remainder of the paper is organized as follows. The next section overviews the related work including object detection, keypoint estimation, graph neural network, and human-object interactions. The proposed Relation Parsing Neural Network (RPNN) made up of two attention-based graphs are described in Section 3. Section 4 reports the experimental results and a conclusion is drawn in Section 5.

2. Related work

Object Detection. Mounts of research efforts had been made on this topic over the past two decades. Especially, the R-CNN based models [22, 4, 5, 6], which are two-stage approaches for addressing object detection task in two steps: candidate RoIs proposal and objects classify, are the leading methods of state-of-the-art results. Based on the shared feature maps, region-wise features can be extracted by RoIpooling operation which enables the feasible computation of higher-order interactions detection task. Our method is built on the Mask R-CNN framework [6].

Keypoint Estimation. Keypoint estimation task draws great attention and it can be mainly divided to two classes,i.e., bottom-up methods and top-down methods. Bottom-up methods [9, 7, 8, 23] directly predict all the keypoints in the image and associate them for individual people. Particularly, [7] exploits global contextual cues from other body parts to associate body parts with individuals by utilizing part affinity fields(PAFs). Top-down methods [24, 25, 26, 27, 6] firstly perform person detection and then detect each person's pose, which obtains great performance under the scenario of multiple people. He et al. [6] extend Faster R-CNN into Mask R-CNN by estimating instance segmentation and keypoints together, which is a suitable backbone for our end-to-end method. To keep robust keypoints estimation performance under the scenario of many people, we adopt Mask R-CNN [6] for object detection and keypoint estimation.

Graph Neural Network. There is an increasing interest in extending neural networks to graph models in recent years. We divide them to two categories. The first category is spectral approach working with a spectral representation of graphs, which is applied successfully for node classification. Specifically, [28] proposes a simplified and efficient method by operating in a 1-step neighborhood around each node. The second category is non-

spectral approach [29, 30] which defines convolutions directly on graph to aggregate spatially close neighbors. [31] proposes Message Passing Neural Network model which presents powerful representation of graph convolutional layers. Moreover, a trivially scalable method which uses attention mechanism is proposed in [21]. In this work, we extend [21] into HOI detection task to automatically perform relationship inferring and node refinement, which generates simple but powerful feature representations.

Human-Object Interaction. Detecting Human-Object Interaction (HOI) concentrates on visual relationship but ultimately is essential for machine to concretely understand human activity. Gupta and Malik [13] contribute to construct V-COCO dataset from COCO [32] dataset. [15] combines human-object region pairs with spatial relationship features for detection and expands HICO [33] dataset for human-object interaction detection in HICO-DET dataset. Gkioxari et al. [16] propose to estimate density map of action target location based on detected human appearance, which solves HOI detection task elegantly. [18] applies an attention mechanism to capture the correlation between body-parts for HOI recognition task. In [14], instance-centric attention is proposed to highlight different regions of image for detecting human-object interactions. Qi et al. [17] introduce a learnable graph structure neural network to HOI recognition task, and extend the network into dynamic settings.

Summary Comparing with the above methods, our method differs mainly in the following aspects. First, instead of extending the learned graph model to obtain powerful representations based on coarse human/objects appearance features in [17], we introduce detailed body parts features, and our model combines a graph structure for feature refinement. Second, unlike [14] proposing image attention based on instance centric, we think objects and body-parts are the most interesting regions that need us to pay attention to. Therefore, based on detected body parts and objects, we explicitly introduce Object-Bodypart Attention mechanism and Human-Bodypart Attention mechanism to focus interesting objects and body parts regions. Third, [18] models the pairwise relationships between body parts. While we focus modeling the relation of object-body part pair as well as human-body part pair. To the best of our knowledge, this is the first work to focus on the pair-wise correlations between body parts and objects in HOI detection.

3. Relation Parsing Neural Network (RPNN)

Human-object interaction (HOI) is an important topic of computer vision. In the past, it is usually modeled by human-object graphs. In fact, the interaction between human and objects is through body parts. In the paper, a novel deep neural network, i.e., Relation Parsing Neural Network (RPNN), is proposed for HOI detection. In particular, the RPNN is made up of two attention based graphs, i.e., Human-Bodypart Graph and Object-Bodypart Graph by an additional component, i.e., body parts.

The overview of the proposed RPNN architecture is depicted in Fig. 2. Given an input image, a Mask R-CNN [6] is exploited to detect all the human/object bounding boxes b_h, b_o and human keypoints kp_h. Given the human h detected, we use human keypoints kp_h to construct the body part boxes $b_{p,h}$. Once b_h, b_o and $b_{p,h}$ are obtained, we extract features from the shared feature map of ResNet-50 C4 by the ROIAlign operation [6] and feed the features to ResNet-50 C5 in order to obtain f_h, f_o and $f_{p,h}$. We denotes $\mathcal{O}$ and $\mathcal{P}_h$ as the sets of the f_o and the $f_{p,h}$ of human h, respectively. In addition, the scene feature f_s is obtained by an adaptive average pooling (14×14), and then the shared feature map of ResNet-50 C4 is fed to ResNet-50 C5.

For HOI detection of each person h, Object-Bodypart Graph $\mathcal{G}_h^O$ is constructed by body parts $\mathcal{P}_h$ and objects $\mathcal{O}$. In most cases, a body part interacts with only one object. Under this observation, given a complete Object-Bodypart Graph $\mathcal{G}_h^O$ for each person h, the adjacency matrix $\mathcal{E}_h^O$ is estimated automatically to measure the relationship between $\mathcal{P}_h$ and $\mathcal{O}$. Specifically, an Object-Bodypart attention is introduced to softly discover the most related object-bodypart for every body part p. Then, the object/body part nodes are refined by message passing, and the updated graph is called as Refined Object-Bodypart Graph $\mathcal{G}_h^{O'}$. The refined object nodes $f_o^{'}$ are used to predict the probabilities $s_{h,o}^a$ of diverse actions for the object o and the human h.

The refined body part nodes $\mathcal{P}_h^{'}$ send bodypart-related action contexts to Human-Bodypart Graph $\mathcal{G}_h^H$ by initializing the body part nodes in $\mathcal{G}_h^H$. In addition, we exact the feature from the whole image, and add it to the Human-Bodypart Graph as a scene node f_s for human node refinement. The relationships between human and other nodes are inferred as adjacency matrix $\mathcal{E}_h^H$ by the Human-Bodypart attention mechanism. Furthermore, the Message Passing gathers all features of the node in $\mathcal{G}_h^H$ to refine human node $f_h^{'}$. After refining all the nodes, the graph is called as Refined Human-Bodypart Graph $\mathcal{G}_h^{H'}$, and the refined human node $f_h^{'}$ further predicts the action s_h^a and density estimation u_h^a of target object location for given action a.

Following the [16], HOI scores $S_{h,o}^a$ are calculated by

$$S_{h,o}^a = s_h \cdot s_o \cdot s_h^a \cdot g_{h,o}^a \tag{1}$$

where s_h and s_o are the classification scores from the detected human box b_h and an detected object box b_o. Here, $g_{h,o}^a$ measures the compatibility between the object box b_o and the predict target object location u_h^a, and is computed by

$$g_{h,o}^a = \exp(\|b_{o|h} - u_h^a\|^2 / 2\sigma^2), \tag{2}$$

$$b_{o|h} = \left\{ \frac{x_o - x_h}{w_h}, \frac{y_o - y_h}{h_h}, \frac{w_o}{w_h}, \frac{h_o}{h_h} \right\} \tag{3}$$

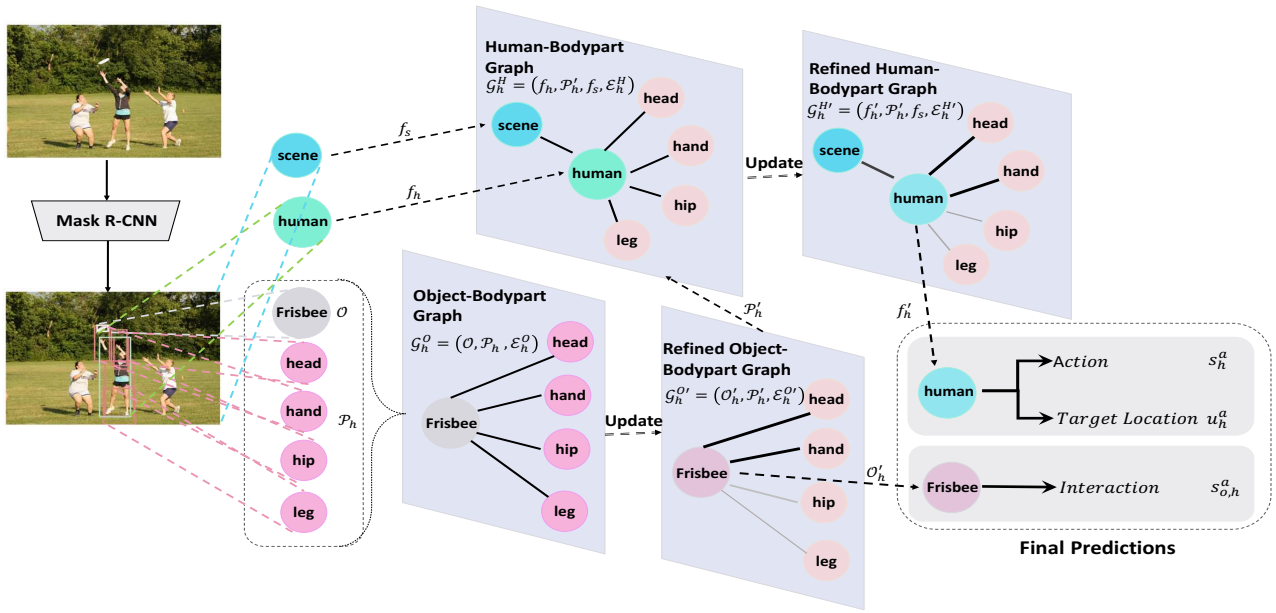

Figure 2. An overview of our Model Architecture for HOI detection (better viewed in color). The update procedure of graph is composed of relation inferring and node refinement (relation is presented by the line thickness between nodes, and the node refinement is illustrated by color change). See Section 3 for detailed explanation.

where $b_{o|h}$ is the encoding of object box b_o in coordinates related to the human box b_h, and we set the hyperparameter σ to 0.3.

The implementation of Object-Bodypart and Human-Bodypart graphs is given in the next two subsections, respectively with the correpsonding attention schemes.

3.1. Object-Bodypart Graph

As discussed in Section 1, the object-bodyparts contain abundant knowledge for the HOI task. Assume that every object can interact with multiple people. Namely, one object may have different action relationships with different people. Hence, one Object-Bodypart Graph is constructed for each person h. The graph is composed of all body part nodes $\mathcal{P}_h$ and all detected object nodes $\mathcal{O}$. Each body part node is connected to all the object nodes. Note that there are no internal links for object nodes or body part nodes. For a body part, assume that there are interaction connections (0/1) with all the peripheral objects. It is difficult to calculate the interaction connections directly without supervised relation labels of an object o and body part p pair. To solve this problem, an attention mechanism called Object-Bodypart Attention, is introduced to implicitly learn the soft relation probability $\alpha_{o,p}$ between an object o and one body part p of a human h. In addition, $\alpha_{o,p}$ can dynamically update the adjacency matrix $\mathcal{E}_h^{O'}$, and controls the messages passing through edges between nodes.

Object-Bodypart Attention To estimate the soft relation probability, a multi-layer perceptron $R_O(f_o, f_p)$ is exploited, where the input is a concatenated feature by connecting object and body part features and the output is evidence score $e_{o,p}$ for each object-bodypart denoted as

$$e_{o,p} = R_O(f_o, f_p). \tag{4}$$

Here, the evidence score indicates the relationship between an object o and a certain body part p. Empirically, a body part usually interacts with one object. Therefore, the soft relation probability or attention score $\alpha_{o,p}$ can be calculated by a softmax function as follows,

$$\alpha_{o,p} = \frac{\exp(e_{o,p})}{\sum_{n=1}^{N} exp(e_{n,p})}, \tag{5}$$

where N is the number of object nodes in an image.

Message Passing After obtaining $\alpha_{o,p}$, a soft adjacency matrix $\mathcal{E}_h^{O'}$ can be updated to refine object nodes by aggregating features of other body part nodes. Particularly, a linear combination of the node features is utilized to generate refined object features by

$$f_o^{'} = f_o + \sum_{p=1}^{N_h} \alpha_{o,p} \cdot f_p, \tag{6}$$

where N_h is the number of body parts of the person h. Accordingly, the refined body part features in Object-Bodypart Graph can be calculated by

$$f_p^{'} = f_p + \sum_{o=1}^{N} \alpha_{p,o} \cdot f_o. \tag{7}$$

Here, the adjacency matrix is assumed as a symmetric one.

3.2. Human-Bodypart Graph

The Human-Bodypart Graph contains a human node f_h and body part nodes $\mathcal{P}_h$. Every body part node is linked to the human node. The features of body part nodes are initialized by the corresponding refined body part features obtained from the refined Object-Bodypart Graph $\mathcal{G}_h^{O'}$. This is defined as a *action passing* mechanism.

The background (or scene feature f_s) often contains prior knowledge for human action detection. For instance, people always ski in the snow and surf in sea. Instead of directly concatenating scene feature with human feature for action prediction through neural networks, a human-bodypart attention scheme is introduced to infer the relation between human f_h and the scene f_s.

Human-Bodypart Attention Similar to Object-Bodypart Graph, paired node features are concatenated and fed into a multi-layer perceptron R_H to obtain relation scores for human-bodypart $\beta_{h,p}$ and human-background $\beta_{h,s}$ respectively by

$$\beta_{h,p} = \sigma(R_h(f_h, f_p^{'})), \tag{8}$$

$$\beta_{h,s} = \sigma(R_h(f_h, f_s)). \tag{9}$$

Note that we use the refined body part feature $f_p^{'}$ from the refined Object-Bodypart Graph $\mathcal{G}_h^{O'}$ to initialize the Human-Bodypart Graph $\mathcal{G}_h^{H}$. Since human can simultaneously perform multiple actions with several body parts, a sigmoid function $\sigma(\cdot)$ is utilized to normalize the relation scores to $(0, 1)$.

Message Passing To obtain a more discriminative human feature representation, we refine the human node by message passing in Human-Bodypart Graph $\mathcal{G}_h^{H'}$. Once adjacency matrix $\mathcal{E}_h^{H'}$ is obtained by calculating β, human node is refined by computing a linear combination of the body part features as

$$f_h^{'} = f_h + \sum_{p=1}^{N_h} \beta_{h,p} \cdot f_p^{'} + \beta_{h,s} \cdot f_s. \tag{10}$$

3.3. Training and Inference

Multi-task Training A person usually performs multiple actions as multi-label classification problem. Here, we apply binary sigmoid classifiers for each action category, and minimize the binary cross entropy losses between scores s_h^a, $s_{h,o}^a$ and their ground-truth action labels. In addition, we minimize the smooth L1 loss [4] between u_h^a and $b_{o|h}$ for the ground-truth triplet $\langle h, a, o\rangle$. Our overall loss is summed over all losses with weight of one, except for the loss term for s_h^a with a weight of two.

Inference Following the cascaded inference algorithm in [16], we firstly discover an object with high confidence. The high-confidence object usually has a high-scoring action, and also is close to the predicted target location. It can be estimated by

$$b_{o^*} = \arg\max_{b_o} s_o \cdot s_{h,o}^a \cdot g_{h,o}^a. \tag{11}$$

After selecting optimal b_{o^*} for each person h and action a pair, we obtain the triplets $\langle human, verb, object\rangle$ to compute the score $S_{h,o}^a$ as final output of our model.

4. Experiments

4.1. Setting

Dataset To verify the effectiveness of our method, we conduct experiments on two HOI benckmark datasets, i.e. V-COCO [13] and HICO-DET [15] datasets. V-COCO (Verbs in COCO) dataset is a subset of COCO [32] that contains abundant annotations for studying HOI detection. V-COCO includes 10,346 images and 16,199 human instances. The trainval set consists of ∼5,000 images of ∼8,000 person instances, and the test set includes ∼4,900 images. Each instance is annotated with 26 different action classes. Note that there exist three actions that are annotated with two types of targets, i.e., instrument and direct object. HICO-DET is the subset of HICO [33]"Humans Interacting with Common Objects". HICO is the largest dataset for HOI detection at present. It contains 80 object categories that are the same as COCO [32] and includes 600 HOI categories over 117 action classes. HICO-DET provides 47,776 images with more than 150,000 annotated instances, where 38,118 images are included in the training set, and 9,658 images in the test set.

Evaluation Metrics Since the main purpose of the HOI detection is detecting triplet $\langle human, verb, object\rangle$, mean Average Precision (mAP) [13] is used to evaluate the proposed model for HOI detection on V-COCO and HICO-DET datasets. A detected triplet is considered as a true positive if both the predicted human box and target object box have Intersection over Union (IoU) greater than or equal to 0.5 of the corresponding ground-truth boxes.

Implementation Details First, the Detectron [34] of Mask R-CNN [6] is trained on COCO training dataset for both object detection and keypoint estimation. The boxes with scores higher than 0.4 will be kept when the human and object bounding boxes are obtained.

For body part generation, the detected keypoints with scores higher than 0.001 and the area of corresponding human boxes which larger than 10,000.0 are selected. In this work, we define four body part regions, i.e., head, hand, hip and leg. The detail keypoint grouping policy for each body part is introduced as follows. Head part consists of "nose, left/right eye, left/right ear". Hand part includes six parts, i.e., "left/right hand, left/right wrist, left/right elbow". Hip

part contains "left/right hip, left/right knee". Leg part includes "left/right ankle". For every person h, the bounding box of body part $b_p = \{x^1_{h,p}, y^1_{h,p}, x^2_{h,p}, y^2_{h,p}\}$ is calculated by

$$\begin{aligned} x^1_{h,p} &= \min_{kp \in \mathcal{KP}_{h,p}} x_{kp} - \tau_p \cdot w_h, \\ y^1_{h,p} &= \min_{kp \in \mathcal{KP}_{h,p}} y_{kp} - \tau_p \cdot h_h, \\ x^2_{h,p} &= \max_{kp \in \mathcal{KP}_{h,p}} x_{kp} + \tau_p \cdot w_h, \\ y^2_{h,p} &= \max_{kp \in \mathcal{KP}_{h,p}} y_{kp} + \tau_p \cdot h_h, \end{aligned}$$

where $\mathcal{KP}_{h,p}$ denotes the keypoint group of body part p, (w_i, h_i) is the size of detected human bounding box, and τ_b dynamically controls the body part size relevant to human size. As the hand keypoint is not detected, we calculate the coordinate of hand by $h_{(x,y)} = w_{(x,y)} - 0.5 \cdot (w_{(x,y)} - e_{(x,y)})$, where $h_{(x,y)}, w_{(x,y)}, e_{(x,y)}$ denote the 2D coordinate of hand, wrist and elbow, respectively.

Note that HICO-DET dataset does not have keypoint annotations. The Mask R-CNN is applied to directly acquire the keypoints for training dataset to generate body parts. Moreover, the hyperparameter τ is set to be $\{0.2, 0.2, 0.25, 0.25\}$ to control the size of $\{head, hand, hig, leg\}$, respectively. As keypoints may be detected in a small area due to an error detection or other reasons, these parts usually cannot capture the enough information. Accordingly, τ guarantees an enough information of body part.

In the experiments, a pre-trained Mask R-CNN [6] with ResNet-50 [1] is used as the backbone. The functions R_O and R_H consist of $FC(4096 \times 256)$ and $FC(256 \times 1)$, and the leakyReLU is applied, non-linearity with negative input slope $\alpha = 0.2$ in the middle. The refined features f'_h and f'_o are fed to one 2048-d fully connected layer for final prediction (i.e. s^a_h, u^a_h and $s^a_{h,o}$). The initial learning rate is set to be 1e-3 and decayed every 10 epochs by 0.1. The SGD optimizer with 0.9 momentum is applied, and the weight decay is set to be 1e-5. V-COCO and HICO-DET datasets comply with the same training procedure.

4.2. Results

Table 1 reports the performance on V-COCO test dataset and its comparison with other state-of-the-art methods. By incorporating Object-Bodypart Graph and Human-Bodypart Graph, the proposed RPNN gets 47.53 mAP on V-COCO test dataset, which achieves the best performance. As shown in Tab. 2, followinng the evaluation protocol [15], our model is evaluated over three different HOI category sets, i.e., all 600 HOI categories in HICO (Full), 138 HOI categories with less than 10 training instances (Rare), and 462 HOI categories with 10 or more training instances. The proposed RPNN network also achieves stable and superior performance on the HICO-DET test dataset.

Table 1. Performance comparison with the state-of-the-art methods on the V-COCO dataset.

Methods	Feature Backbone	mAP$_{role}$ Scenario 2 $\langle human, verb, object \rangle$
Gupta et al. [13] (implemented by [16])	ResNet-50-FPN	31.8
InteractNet [16]	ResNet-50-FPN	40.0
GPNN [17]	Deformable ConvNets [35]	44.0
iCAN [14]	ResNet-50	45.3
RPNN	ResNet-50	**47.53**

Table 2. Performance comparison with the state-of-the-art methods on the HICO-DET dataset.

Methods	Feature Backbone	Default		
		Full	Rare	Non Rare
Shen et al. [36]	VGG-19	6.46	4.24	7.12
HO-RCNN [15]	CaffeNet	7.81	5.37	8.54
InteractNet [16]	ResNet-50-FPN	9.94	7.16	10.77
GPNN [17]	Deformable ConvNets [35]	13.11	9.34	14.23
iCAN [14]	ResNet-50	14.84	10.45	16.15
RPNN	ResNet-50	**17.35**	**12.78**	**18.71**

Figure 3 shows the visualization examples of the predicted body part attentions for objects on the V-COCO test dataset. For a clear view, we present the most related visual result of one body part, and the prediction of attention score is displayed on the top of object box. As one can see, the head pays most attention to looking the skateboarding, the leg on snowboarding the snowboard, the hand on catching the Frisbee.

Besides, Fig. 4 shows the visualization examples of the predicted human attention for body parts on the V-COCO test dataset, where we only present the visual results of one human, and the score of attention is displayed on the top of the body part boxes. One can clearly see that the human pay more attention to the body parts which interact with objects. Furthermore, our model implicitly learns to always pay more attention to head part.

Ablation Studies To evaluate the effectiveness of different components in the proposed model, seven ablation study experiments are conducted on the V-COCO dataset.

For a comprehensive evaluation on the model components, we also evaluate performances in Scenario 1 from [13]. In this scenario, the corresponding role prediction must be empty with missing role annotations in test cases. This scenario is suitable for missing roles due to occlusion. Furthermore, the RPNN without two graphs is similar with InteractNet [16], which is set as the baseline in our ablation studies. The baseline result obtains the similar performance as recorded in [16]. The comparable results are listed in Tab. 3- 5.

with vs. without action passing. A variant of the proposed method is studied without action passing. Instead of using refined body part features from refined Object-Bodypart Graph, the Human-Bodypart Graph directly uses body part features for initialization. In our model, human feature can potentially capture detailed context of object-bodypart in refined body part features. Tab. 3 reveals that

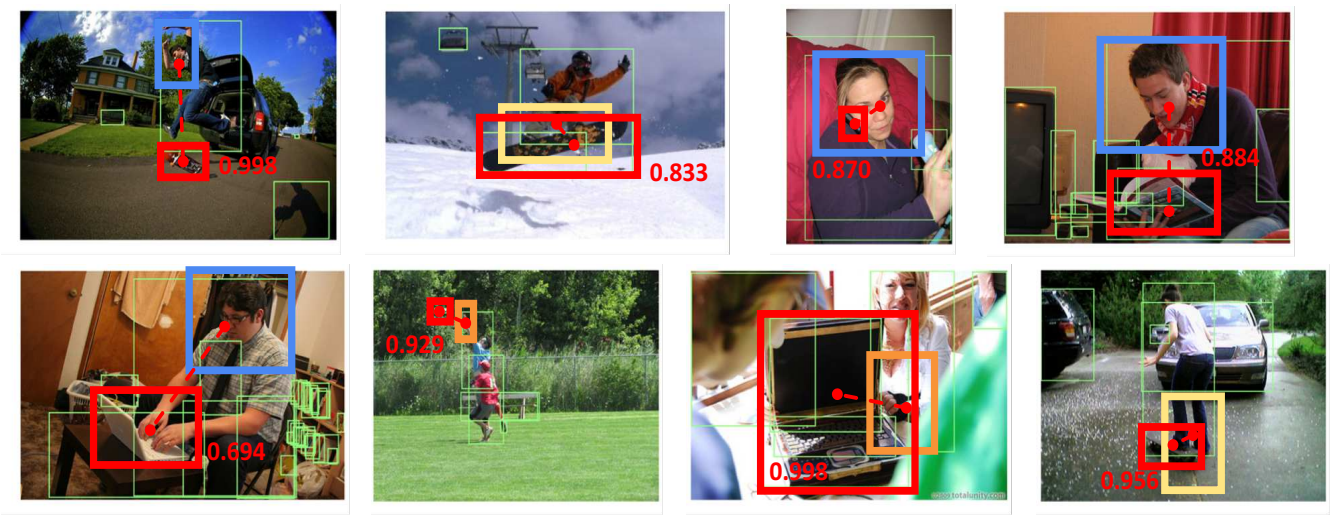

Figure 3. The body part is able to pay more attention to the most related object (marked in red) on V-COCO test dataset, and only one body part of a person is presented for a clear display. The object-bodyparts with the highest score of attention are shown, where the head part is marked in blue, the hand part in orange, the hip part in pink, the leg part in yellow and other detected object boxes in green.

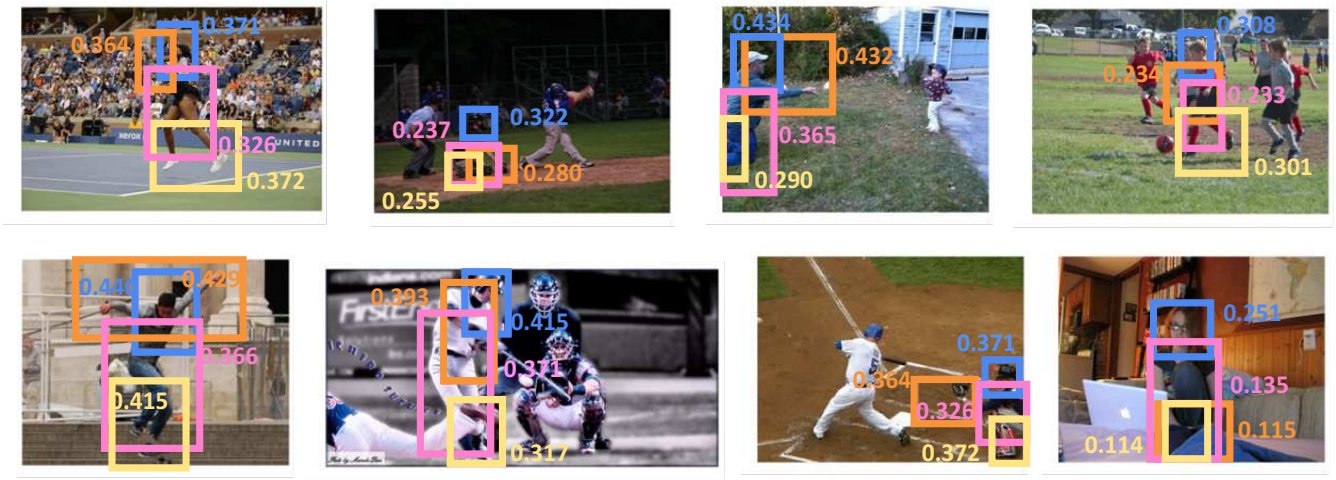

Figure 4. The human is able to pay more attention to the more related body parts on V-COCO test dataset. The human-bodyparts with the attention score are displayed and the different colors of boxes have the same meaning as in shown in Fig. 3.

AP_{role} drops 2.57% and 3.52% on Scenario 1 and Scenario 2 respectively.

with vs. without object-bodypart graph. Object-Bodypart Graph is the key component. It not only refines object nodes but also generates refined body part features for Human-Bodypart Graph initialization. To evaluate this influence, a variant of the proposed model is implemented. The performances drop obviously 4.83% and 5.85% in Scenario 1 and Scenario 2 as shown in Tab. 3.

with vs. without human-bodypart graph. In our model, we use Human-Bodypart Graph as it refines human feature by refined body part features. A variant is implemented without Human-Bodypart Graph. The experiments show a degradation of 4.29% and 5.33% in Scenario 1 and Scenario 2 without Human-Bodypart Graph.

RPNN vs. Different sequences of message passing. Usually, human h might interact with multiple objects, e.g., $\{o_i, o_j, o_k\}$. Hence, f_h might contain all the interaction information with $\{o_i, o_j, o_k\}$. Accordingly, directly adding f_h to the object feature $f_{o,i}$ of the object o_i would bring a redundant noisy of $\{o_j, o_k\}$ to $f_{o,i}$. Therefore, directly using human feature to refine object feature such as "human $\leftrightarrow$ body part $\rightarrow$ object "would cause the performance degradation.

Moreover, due to an imperfect object detection system, many redundant overlappingly detected object boxes are usually detected. Using these overlapping object boxes for refining the human feature would generate an unstable hu-

Table 3. Ablation on the V-COCO test dataset about the architecture design.

Methods	mAP_{role} Scenario 1	mAP_{role} Scenario 2
baseline [16] w/ ResNet-50	-	40.0
baseline [16] (our re-implementation)	30.68	40.23
RPNN w/o action passing	34.11	44.01
RPNN w/o object-bodypart graph	31.85	41.68
RPNN w/o human-bodypart graph	32.39	42.20
human $\leftrightarrow$ body part $\rightarrow$ object	34.26	44.30
bodypart $\leftrightarrow$ object $\rightarrow$ human	33.64	44.24
Human-bodypart-object graph	36.31	46.70
RPNN	**36.68**	**47.53**

Table 4. Ablation on the V-COCO test dataset about the design of attention. human attention: Human-Bodypart Attention; object attention: Object-Bodypart Attention.

Methods	mAP_{role} Scenario 1	AP_{role} Scenario 2
RPNN w/o human attention	33.84	44.59
RPNN w/o object attention	32.73	42.86
RPNN	**36.68**	**47.53**

man feature due to large numerical fluctuations. Hence, directly using object feature for human feature refinement such as "body part $\leftrightarrow$ object $\rightarrow$ human "would decrease the HOI performance.

As a body part only interacts with one object, the RPNN (object$\leftrightarrow$body part$\rightarrow$human) not only uses body part to refine object features, but also uses body part to select the most related object for human feature refinement, which overcomes the problem aforementioned. Therefore, RPNN can obtain the best result as shown in Tab. 3.

RPNN vs. Human-bodypart-object graph. Human-bodypart-object graph contains all nodes. Different from building such a big graph, we use two graphs in order to explicitly adds a prior knowledge to the model. The prior knowledge is that the direct message passing between human and objects is not allowed. As for this big graph, the connection between human and objects should be weakened in order to reduce an information loss. The experimental results support the above discussion as shown in Tab. 3.

with vs. without the attention module. The attention mechanism is also the essential component in our network as it can not only enhance the related context but also suppress the useless ones. Two variants without human-bodypart attention and without object-bodypart attention are carried out respectively. The results shown in Tab. 4 verify that removing human-bodypart attention in Human-Bodypart Graph reduces the performance by 2.84% and 2.94%, and removing object-bodypart attention in Object-Bodypart Graph reduces the performance by 3.95% and 4.67%.

Table 5. Ablation on the V-COCO test dataset about the design of scene features.

Methods	mAP_{role} Scenario 1	AP_{role} Scenario 2
RPNN w/o scene feature	36.25	46.77
RPNN	**36.68**	**47.53**

with vs. without the scene features. The scene feature is used to take the background into consideration. A variant without scene node in Human-Bodypart Graph is further studied to evaluate the effectiveness of the scene feature. The result is shown in Tab. 5, where the performance drops by 0.43% and 0.76% in Scenario 1 and Scenario 2.

From the above ablation studies, we can observe that these components and the sequence of message passing are crucial to the HOI task, since they capture knowledge from different perspectives. Our proposal aggregates these crucial information and thus provides a better result.

5. Conclusion

In the paper, a novel deep neural network is proposed to detect human object interactions (HOI) by introducing "bodypart" together with "human" and "object". In particular, two attention based graphs are constructed to capture the detailed information and parsing relations of object-bodypart and human-bodypart, which are denoted as Object-Bodypart Graph and Human-Bodypart Graph, respectively. The two graphs are combined to construct a Relation Pasing Neural Network (RPNN) for HOI detection by integrating the additional element "bodypart" to the deep neural network. Here, the Object-Bodypart Graph dynamically updates the adjacency matrix to capture the relationships between object and bodyparts based on an object-bodypart attention mechanism, and refines the bodypart/object features by message passing. We perform an action passing scheme by using refined body part nodes to initialize Human-Bodypart Graph, and enrich the human feature representation by message passing. The experiments conducted on HOI benchmark V-COCO and HICO-DET datasets confirm the effectiveness of the proposed RPNN network, which can significantly improve the detection accuracies compared with the state-of-the-art models.

Acknowledgments

This work was supported in part by National Key R&D Program under contract (2017YFA0402600), and in part by Shanghai Fabric Eyes AI Technology Co., Ltd.

References

[1] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[2] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *CVPR*, 2015.

[3] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *AAAI*, 2017.

[4] Ross Girshick. Fast R-CNN. In *ICCV*, 2015.

[5] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *NIPS*, 2015.

[6] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B. Girshick. Mask R-CNN. In *ICCV*, 2017.

[7] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, 2017.

[8] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *ECCV*, 2016.

[9] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. Convolutional pose machines. In *CVPR*, 2016.

[10] Abhinav Gupta, Aniruddha Kembhavi, and Larry S Davis. Observing human-object interactions: Using spatial and functional compatibility for recognition. *TPAMI*, 2009.

[11] Vincent Delaitre, Ivan Laptev, and Josef Sivic. Recognizing human actions in still images: a study of bag-of-features and part-based representations. In *BMVC*, 2010.

[12] Yao Bangpeng and Li Fei-Fei. Modeling mutual context of object and human pose in human-object interaction activities. In *CVPR*, 2010.

[13] Saurabh Gupta and Jitendra Malik. Visual semantic role labeling. *CoRR*, 2015.

[14] Chen Gao, Yuliang Zou, and Jia-Bin Huang. ican: Instance-centric attention network for human-object interaction detection. In *BMVC*, 2018.

[15] Yu-Wei Chao, Yunfan Liu, Xieyang Liu, Huayi Zeng, and Jia Deng. Learning to detect human-object interactions. In *WACV*, 2018.

[16] Georgia Gkioxari, Ross Girshick, Piotr Dollár, and Kaiming He. Detecting and recognizing human-object interactions. In *CVPR*, 2018.

[17] Siyuan Qi, Wenguan Wang, Baoxiong Jia, Jianbing Shen, and Song-Chun Zhu. Learning human-object interactions by graph parsing neural networks. In *ECCV*, 2018.

[18] Hao-Shu Fang, Jinkun Cao, Yu-Wing Tai, and Cewu Lu. Pairwise body-part attention for recognizing human-object interactions. In *ECCV*, 2018.

[19] Ty W Boyer, Josita Maouene, and Nitya Sethuraman. Attention to body-parts varies with visual preference and verb–effector associations. *Cognitive processing*, 2017.

[20] Tony Ro, Ashley Friggel, and Nilli Lavie. Attentional biases for faces and body parts. *Visual Cognition*, 2007.

[21] Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. In *ICLR*, 2018.

[22] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, 2014.

[23] Eldar Insafutdinov, Leonid Pishchulin, Bjoern Andres, Mykhaylo Andriluka, and Bernt Schiele. Deepercut: A deeper, stronger, and faster multi-person pose estimation model. In *ECCV*, 2016.

[24] Alejandro Newell, Zhiao Huang, and Jia Deng. Associative embedding: End-to-end learning for joint detection and grouping. In *NIPS*, 2017.

[25] George Papandreou, Tyler Zhu, Nori Kanazawa, Alexander Toshev, Jonathan Tompson, Chris Bregler, and Kevin Murphy. Towards accurate multi-person pose estimation in the wild. In *CVPR*, 2017.

[26] Fangting Xia, Peng Wang, Xianjie Chen, and Alan L Yuille. Joint multi-person pose estimation and semantic part segmentation. In *CVPR*, 2017.

[27] Haoshu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu. Rmpe: Regional multi-person pose estimation. In *ICCV*, 2017.

[28] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *ICLR*, 2017.

[29] David K Duvenaud, Dougal Maclaurin, Jorge Iparraguirre, Rafael Bombarell, Timothy Hirzel, Alán Aspuru-Guzik, and Ryan P Adams. Convolutional networks on graphs for learning molecular fingerprints. In *NIPS*, 2015.

[30] James Atwood and Don Towsley. Diffusion-convolutional neural networks. In *NIPS*, 2016.

[31] Justin Gilmer, Samuel S Schoenholz, Patrick F Riley, Oriol Vinyals, and George E Dahl. Neural message passing for quantum chemistry. In *ICML*, 2017.

[32] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *ECCV*. 2014.

[33] Yu-Wei Chao, Zhan Wang, Yugeng He, Jiaxuan Wang, and Jia Deng. HICO: A benchmark for recognizing human-object interactions in images. In *ICCV*, 2015.

[34] Ross Girshick, Ilija Radosavovic, Georgia Gkioxari, Piotr Dollár, and Kaiming He. Detectron. `https://github.com/facebookresearch/detectron`, 2018.

[35] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *ICCV*, 2017.

[36] Liyue Shen, Serena Yeung, Judy Hoffman, Greg Mori, and Li Fei-Fei. Scaling human-object interaction recognition through zero-shot learning. In *WACV*, 2018.