Spatial-Aware Multi-Level Parsing Network for Human-Object Interaction

Zhan Su[1], Ruiyun Yu[1]*, Shihao Zou[2], Bingyang Guo[1], Li Cheng[2]

[1] Software College, Northeastern University, Shenyang (China)
[2] University of Alberta, Edmonton (Canada)

* Corresponding author: yury@mail.neu.edu.cn

Received 6 October 2022 | Accepted 16 May 2023 | Published 27 June 2023

unir LA UNIVERSIDAD EN INTERNET

Abstract

Human-Object Interaction (HOI) detection focuses on human-centered visual relationship detection, which is a challenging task due to the complexity and diversity of image content. Unlike most recent HOI detection works that only rely on paired instance-level information in the union range, our proposed Spatial-aware Multi-level Parsing Network (SMPNet) uses a multi-level information detection strategy, including instance-level visual features of detected human-object pair, part-level related features of the human body, and scene-level features extracted by the graph neural network. After fusing the three levels of features, the HOI relationship is predicted. We validate our method on two public datasets, V-COCO and HICO-DET. Compared with prior works, our proposed method achieves the state-of-the-art results on both datasets in terms of mAP_{role}, which demonstrates the effectiveness of our proposed multi-level information detection strategy.

Keywords

Computer Vision, Deep Learning, Graph Neural Network, HOI Detection, Image Understanding.

DOI: 10.9781/ijimai.2023.06.004

I. Introduction

Images are the main form of information obtained by humans. In recent years, basic vision tasks, such as target detection, action recognition and image segmentation, have developed rapidly with the application of deep learning. Research on higher-level image semantics of individual instances, such as human action recognition and pose estimation, has also made significant progress. Human-Object Interaction (HOI) detection, an intersecting area of object detection [1], behavior recognition [2], and visual relationship detection [3], utilizes images as input to detect and locate human-object pairs and predict their interaction categories. Formally, Visual Relationship Detection (VRD) uses ⟨*objectA, predicate, objectB*⟩ to define relational expressions, which involves a combination of interactions of multiple target objects, such as human-human, human-object, object-object, etc. HOI detection only focuses on the interaction between humans and objects, and the predicates are mainly concentrated in the category of verbs, which have significant reference value for the development of behavior recognition. The diversity of the interaction between humans and objects mainly relies on the object category, the human's pose, and the human's relative position with the object. In some cases, even if the object, pose, and relative position are the same, the behavior could be different. For example, putting things back and picking things up are two different behaviors. These situations make the HOI a challenging task [4].

For the HOI detection task on static images, Gupta and Malik [5] first solved this problem. Before their work, most researchers only recognize human actions and bounding boxes in a single or multiple frames. Based on object detection, they associate various semantic instances in the scene, detect the performed actions and recognize the interactions between object instances and humans, which provides a detailed understanding of current activities. Georgia Gkioxari et al. [6] propose a human-centric architecture strategy for detecting all interactions between objects and humans, considering the number of objects in an image is unknown. The mainstream methods in HOI detection tasks [7]–[10] adopt a similar human-centered network model structure. Gao et al. [7] propose the human-centered attention module iCAN to emphasize the contextual information areas in the image related to interaction. The main idea of this module is to utilize the softmax function to transform the fused instance-level appearance features and convolutional features to an attention map, which produces high-level features. Wang et al. [8] improves the iCAN module by embedding the context-aware appearance and attention modules in the "human stream" and "object stream" to extract the appearance and context information in the image. Bansal et al. [9] proposes the spatial priming model to strengthen the relative spatial position between humans and objects. Liao et al. [10] propose a single-stage parallel point detection and matching model, where the point detection branch estimates human points, interaction points and object points, and the point matching branch takes the human and object points with corresponding interaction points as a matching pair. This

Please cite this article as:
Z. Su, R. Yu, S. Zou, B. Guo, L. Cheng. Spatial-Aware Multi-Level Parsing Network for Human-Object Interaction, International Journal of Interactive Multimedia and Artificial Intelligence, vol. 9, no. 2, pp. 39-48, 2025, http://dx.doi.org/10.9781/ijimai.2023.06.004

method screens out many candidates interaction points to solve the large model scale, increasing detection speed. Although these methods have achieved significant improvement in HOI detection, they do not consider the spatial configuration relationship between all objects and humans in the scene at once and do not make full use of the pose and part-level information of human instances.

The HOI of a single interaction pair is related to its spatial positions, the global interaction relationship with other instances, the pose, and specific parts of a human body. In order to obtain more information, we propose a Spatial-aware Multi-level Parsing Network (SMPNet) for HOI detection. Specifically, an image is an input to the backbone of Faster RCNN [11] and CPN [12] to obtain the feature maps and human pose and then fed into the multi-branch deep network to perform relational inference based on the multi-level features of each human-object pair. Fig. 1 shows an example of our relation reasoning.

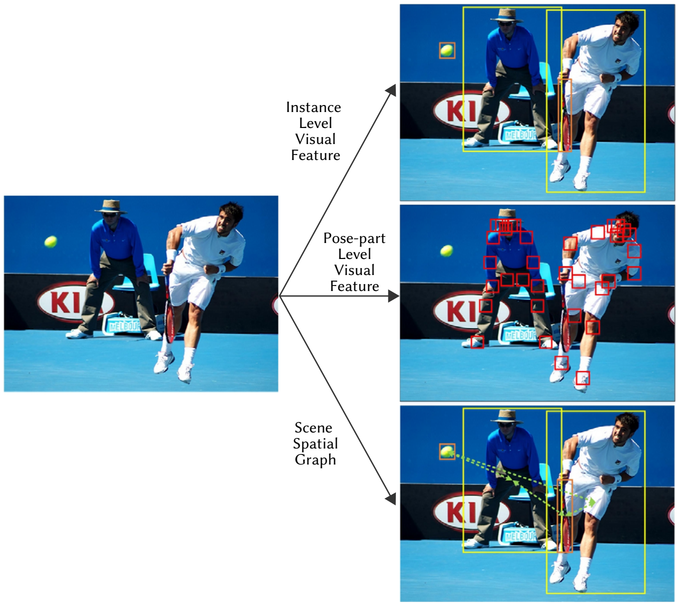

Fig. 1. SMPNet uses three levels of information, including i) the visual features of humans (yellow boxes), objects (orange boxes), and the overall context of an image, ii) the visual features of various parts of the human body (red boxes), iii) the information propagated on the scene spatial graph (green lines).

SMPNet consists of an instance-level and a part-level visual branch to process the instance-level and part-level appearance features, respectively, which are further refined by the spatial attention scores and features from the spatial configuration map. The pose key-points of the target human are obtained through the pose estimator. Then, we embed the encoded pose features into the spatial configuration features within the human bounding box. At the same time, we use the key-points as the center point to obtain the part-level features. Then, the attention weight is extracted to achieve refined instance-level and part-level features. Although some works [7]–[9] have used these spatial configurations directly as classification features, they do not combine pose information with spatial information, and therefore neglect it as a clue to infer part-level human features. Our approach provides attention mechanisms for refining visual features from multiple levels.

In the spatial graph branch, to effectively utilize the spatial configuration information of all instances in the entire scene, we model the scene in the image as a graph, the nodes of which are all the detected humans and objects, and integrate the spatial configuration information to define the propagation on the edges. In this case, each human node can receive lots of messages indicating the existence of other object nodes. If we assume the messages sent from a single object node to all human nodes are the same, then the only variable is the propagation weights that control the information propagation. On this basis, each human node can receive information about the objects' existence and their relative position. After information propagation, node features can effectively integrate each instance's visual and spatial location information in the image and provide additional scene-level clues for interactive detection. Finally, we can predict the HOI categories for each sample based on fused features.

Using the V-COCO [5] dataset and the HICO-DET [13] dataset, we conduct extensive experiments. The observed results reveal that SMPNet surpasses the state-of-the-arts. In general, the following are the major contributions of this research:

- To integrate and extract the correlation between interactive instances in the global environment, we present a special graph neural network design.
- We use the spatial configuration map containing the pose information to obtain the visual appearance features' attention weight and refine the potential multi-level features.
- We use the encoded spatial features combined with the visual appearance features to obtain the message so that the information propagation in the graph neural network process is adjusted according to the interaction pair situation and the sender situation.

The rest of this article is organized as follows. Section II introduces the related work. Section III explains the SMPNet model comprehensively. Extensive experiments are executed in Section IV to demonstrate the efficacy of our method. Lastly, Section V concludes the article.

II. Related Work

A. Attention Mechanism

The attention mechanism mainly utilizes human vision to quickly scan the global image and then obtain the target area in the image that needs attention. This idea is to imitate the special brain signal processing mechanism of human vision and appropriately invest more attention in the target area to obtain more detailed information and suppress other useless information. Attention mechanism (AM) mainly builds an attention matrix to make the deep neural network pay attention to the key features in the image during the training process to avoid the impact of non-key features. The attention mechanism is first applied in machine vision, and its main function is to make the areas that need to be focused on in the data to get more attention [14]–[19], Bahdanauu et al. [18] apply the attention mechanism in the machine translation task. Their work is further verified that the attention mechanism can effectively reflect the relationship between features, promoting deep neural networks combined with attention mechanism research. Subsequently, VaSWanl et al. [19] introduce the attention mechanism into the sentence modeling task and use a two-dimensional matrix to represent sentence information, thereby obtaining a feature representation of richer semantic information. Later, researchers introduce the attention mechanism into image processing, such as Gao et al. [7]. Our task focuses on part-level and instance-level attention for HOIs detection.

B. Object Detection

Traditional target detection algorithms can be divided into target instance detection and traditional target class detection. The former considers that the objects in the image are irrelevant except for the specific target of interest. The detection target instance usually uses the template and image stability features to obtain correspondence between the objects in the scene. The latter is based on the selected features and classifiers using HOG [20] features, support vector machine [21] and AdaBoost [22] algorithm framework and other methods to detect a limited number of classes.

Alex et al. propose the Alexnet convolutional neural network model and improve it significantly. Compared with traditional algorithms that manually extract features, deep neural networks have considerably improved in nature. Since then, deep neural networks' powerful autonomous learning and expression capabilities have replaced traditional feature extraction methods. Target detection methods based on deep learning include two-stage target detection algorithms and one-stage target detection algorithms. RCNN [23], Fast RCNN [24], Faster RCNN [11], etc., are common two-stage target detection algorithms, and YOLO [25], SSD [26], etc., are common one-stage target detection algorithms. The two-stage target detection algorithm uses a convolutional neural network to classify the generated candidate frame samples. The one-stage target detection algorithm is different from the two-stage target detection algorithm. It does not need to generate a candidate frame but directly converts the problem of target frame positioning into a regression problem. Additionally, we utilize the Faster RCNN with ResNet-50-FPN as a region proposal network and extend it to interaction proposals that predict if a human-object pair is interacting.

C. Visual Relationship Detection Technology

Image understanding can identify the relationship between objects in an image and form a comprehensive language description. It is one of the widest applications in image processing. In general, an image usually contains multiple interacting objects. Recognizing a single object is not enough to better understand these images. The relationship between objects also contains very important information. And sometimes, this relationship determines the semantic information represented by this image. This has led to the diverse relationships in the image becoming the focus and difficulty of image understanding.

Visual relationship detection technology is mainly divided into visual relationship detection based on scene graphs and visual relationship detection based on visual features. The former utilizes scene graphs to understand images' high dimensional semantic meaning as a problem of obtaining a directed graph structure. Johnson et al. [27] first propose the concept of Scene Graphs, which can more accurately understand the semantic information of images. Jianwei Yang et al. propose the Graph-R CNN framework [28], which utilizes graph convolution based on the directed graph structure to identify the relationship between objects. The relation proposal network (RePN) model they proposed effectively solves the problem that the connection between two objects increases with the number of objects squared. In addition, they also propose an attention graph convolutional network (aGCN), which can efficiently obtain objects' interrelationships between objects instances. The latter's representative is the visual transformation embedding network (VTransE) [29] proposed by Hanwang Zhang et al. for visual relationship detection. The characteristic of the network is to put the target in the designed low-dimensional relational space and utilize simple vector transformation in this space to express the relation between objects. For example, the subject&predicate are approximately equal to the object. At he same time, they utilize a novel feature extraction layer, which enables the transfer of target relationship knowledge in the form of full convolution and trains in a simple forward or backward. In our research, our focus is on HOI detection, which is a human-centric problem, to detect action interactions between humans and objects.

D. HOI Detection

Based on the human-centric network model and the human-object region convolutional neural network (HO-RCNN), Chen Gao et al. propose an end-to-end trainable attention network iCAN [7]. First, the network utilizes humans as the center of interaction to obtain the position probability map of the object interacting with the human. Through the position probability map, they clarify the relationship pair. Based on the iCAN model, many variants using mixed approaches [8], [30]–[32] have been proposed. Wang et al. [8] proposed a contextual attention model, which can adaptively learn the context information of instances, allowing the network to focus on important semantic areas. Xu et al. [30] constructed a priori-knowledge graph based on the annotation of the HOI dataset and the external visual relationship dataset and modeled the semantic relationship between verbs and objects to alleviate the long-tail distribution problem. Wang et al. [31] started from the region proposal module, proposed an HO-RPN network suitable for HOI detection, and introduced an external word embedding in the object classification to achieve Zero-Shot Learning. Bansal et al. [32] designed the functional generalization module, which uses the word embedding of human and object as a feature to endow the model with zero-shot learning capabilities. Song Gao et al. [33] improve the accuracy of the existing model by optimizing the loss function and training details to improve the performance of the HOI detection task. However, those works only take instance-level features of humans and objects but do not utilize the features in each part of the human, and other scene instances with spatial information, which provides more detailed information for the HOI task. We propose a model that captures multi-level information from input feature maps.

III. Spatial-Aware Multi-Level Parsing Network

This section introduces our proposed SMPNet for detecting human-object interaction. As shown in Fig. 2, we integrate the features and values of three individual branches to perform interaction prediction: the first is utilized to analyze human part-level visual features, the second is applied to analyze instance-level visual features, and the third is used to extract interactive features between all instances in the scene as scene-level features. The spatial configuration is an important reference feature that indicates whether an object interacts with a human. For example, when the object and the human proposals are close in space, or even the object proposal is close to the spatial position of any pose key-point, they usually have direct visual interaction. In contrast, there is usually no direct visual interaction when two proposals are spatially far apart. According to this property, introducing a spatial configuration map to generate weights can better refine the weights of instance-level and part-level features. Then, the spatial features between the interaction pairs are coded together with the appearance features and used as the input to the spatial graph branch, and the interaction features are obtained through message passing of the graph neural network. This branch considers multiple groups of interactions involving multiple humans and multiple objects in the scene. These interaction features can provide additional reference features for detecting human-object pairs. For example, the content in the image is a concert, and the information that other humans are playing some musical instruments in the image can provide reference features for the currently detected human-object pair playing the piano.

We extract the appearance features of instances and context (Sec. A) based on the work of Chen et al. [7]. Following that, we explain the fundamental method that incorporates the spatial configuration's attention mechanism, including the pose information into this branch network. We use human pose cues to improve the local semantics on the spatial configuration map and extract attention scores (Sec. B). Then, we utilize these scores to enhance the visual appearance features of part-level. Next, we obtain relevant features within instances based on a bipartite scene graph (Sec. C). At last, we describe the fusion of instance-level visual features, human part-level visual features, and scene-level instance-related features (Sec. D).

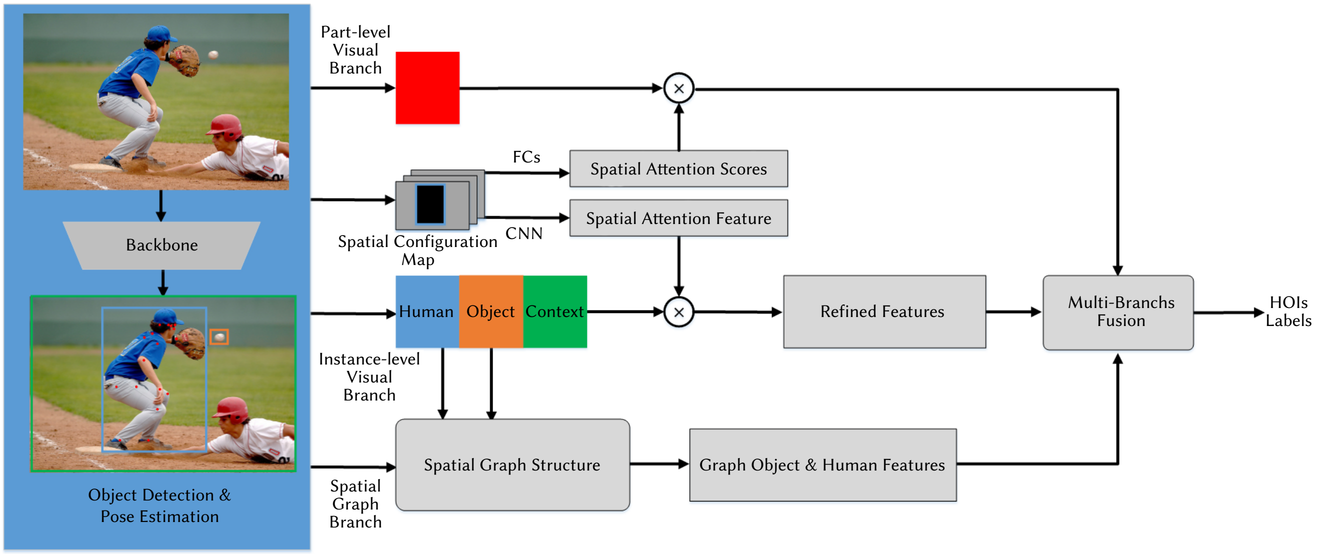

Fig. 2. Overview of our framework: For an interaction pair (a human and an object), the "backbone module" aims to prepare convolutional feature maps and the ROIs for three parallel branch networks. Rounded rectangles are operations, and ⊕ is element-wise multiplication.

A. Instance-Level Visual Branch

We construct a branch network based on the method in [7] to extract instance-level visual appearance features, including human area, object area, and scene context. Detailed information is shown in Fig. 3.

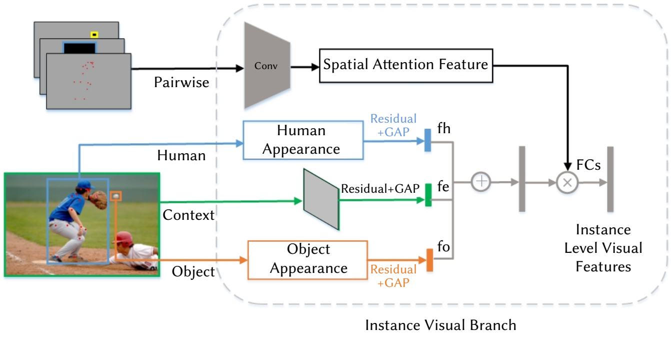

Fig. 3. Structure of the instance level visual branch. The branch contains human, object, context, and pairwise spatial streams. Here ⊕ is the concatenation process, ⊗ represents element-wise multiplication, GAP is global average pooling, Residual denotes residual block [34], and FCs denotes two fully connected layers.

This branch focuses on extracting the visual appearance features of human-object pairs. Referring to the multiple variants of the iCAN model using mixed approaches [8],[30]–[32], we designed the instance-level visual branch that includes four proven effective feature streams: object, human, context, and pairwise streams. Compared to [7], we utilize RoIAlign rather than ROI pooling and adjust the dimensions of each part. RoIAlign is a method proposed in Mask-RCNN [35] to aggregate and output specified size features in regions of different sizes on the feature map. It uses the bilinear interpolation method to obtain the feature values on the pixels whose coordinates are floating-point numbers, thus transforming the whole process of feature aggregation into a continuous operation. We use RoIAlign on the human and object regions to extract features following object detection. This operation is followed by a residual block (Res) [34] and global average pooling (GAP) to extract visual feature vectors of objects, humans and context.

In contrast to the late fusion approach in [7], we apply the early fusion approach to process the features of the human, object, and context features as f_h, f_o, f_e, which is to concatenate all the features and project it to obtain the instance-level visual appearance feature as equation (1):

$$f_{ivis} = W_{ivis}(f_h \oplus f_o \oplus f_e) \tag{1}$$

where ⊕ denotes concatenation operation, W_{ivis} is the projection matrix which is realized through a fully connected layer, f_{ivis} is a feature vector of size D.

To focus on learning the spatial interaction mode between humans and objects, the output of the pairwise stream is the attention feature. This attention feature is used to refine the visual features of humans, objects, and the context obtained in the other three streams, as shown in Fig. 3.

We use the two binary masks of humans and objects proposal in the image as clues to capture the instance-level spatial configuration, similar to [7], [8], [36]. In detail, according to the given human proposal x_h and the object proposal x_o, we generate two binary images. These binary maps have zeros everywhere except for the region defined by the human and object proposal x_h and x_o of each map, respectively. At the same time, to match the pose-parts' visual features, we encode the pose key-points extracted in the backbone stage into a map according to the skeleton configuration of the coco dataset following the work of Yong-Lu Li [37]. In this map, the key-points are connected by lines with diverse gray values between 0.1 and 0.9, indicating the corresponding parts of the pose, and the values of the remaining areas are all 0. After that, the two kinds of maps are rescaled to the size of $M \times M$ and connected to generate a 3-channel binary scene spatial configuration map M_{hop}.

Referring to the work of [7], [13], we utilize two convolutional layers to parse the scene spatial configuration map. The following equation (2) is the GAP and full connection operation:

$$a_{hop} = W_{hop}(GAP(Conv(M_{hop}))) \tag{2}$$

where W_{hop} denotes a fully connected layer and a_{hop} is the attention feature vector of the same size as the visual feature vector f_{ivis} obtained in the previous branch.

On this basis, we utilize a_{hop} to refine the visual feature vector f_{ivis} through the vector multiplication operation to obtain the output f_a of this branch. The formula (3) is as follows:

$$f_a = f_{ivis} \otimes a_{hop} \tag{3}$$

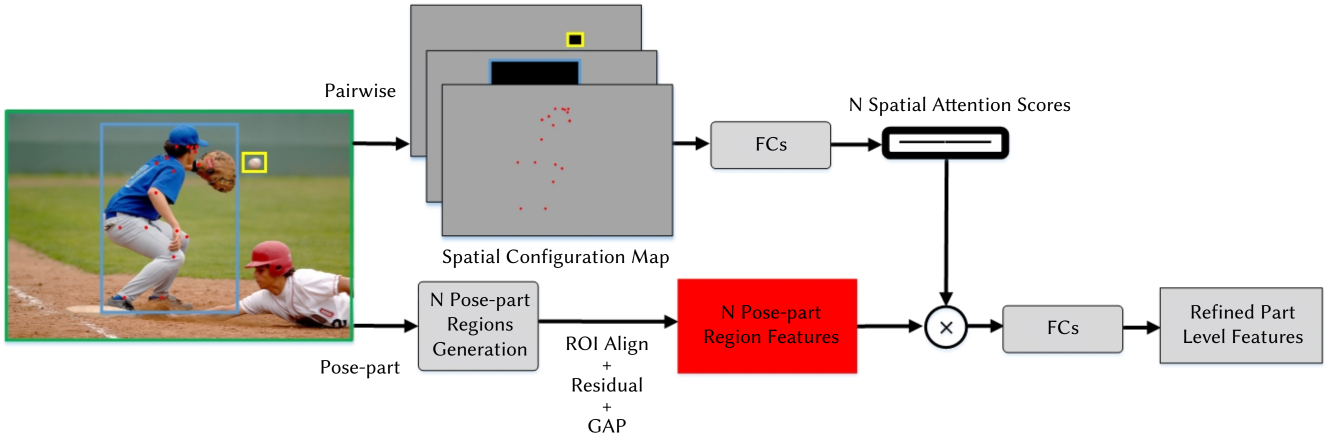
Pairwise
Spatial Configuration Map
FCs
N Spatial Attention Scores
Pose-part
N Pose-part Regions Generation
ROI Align + Residual + GAP
N Pose-part Region Features
×
FCs
Refined Part Level Features

Fig. 4. Structure of the part-level visual branch. The overall module includes pose-part and pairwise spatial streams. Here ⊕ denotes concatenation operation, ⊗ is element-wise multiplication, Residual denotes block [34], GAP is global average pooling, and FC denotes fully connected layer.

where ⊗ is the element-wise multiplication, and f_a is the refined feature vector of size *D*.

B. Part-Level Visual Branch

As shown in Fig. 4, this branch focuses on extracting the visual features of the part-level. First, we extract the part area of each key-point representing *N* human body parts. From the backbone stage, we obtain all the pose key-points $k(p)$ of the human *h*. Every pose point serves as the center for the generation of an area $R_{pk} = \{h_{pk}, w_{pk}, x_{pk}, y_{pk}\}$. Furthermore, during the process of generating the area, it is stipulated that it will not exceed the range of the image. The computation procedure is as equation (4):

$$h_{pk} = w_{pk} = \lceil \delta\sqrt{h_{human} * w_{human}} \rceil \tag{4}$$

where w_{human} and h_{human} indicate the size parameters of the human bounding box, the notation ⌈⌉ denotes the rounding-up process, and δ indicates the scaling value that is adjusted to 0.1 based on an experimental evaluation.

Then, based on the resulting pose-part areas $R_{parts} = \{R_{p1}, \dots, R_{pN}\}$, we utilize the RoIAlign algorithm and GAP operation extracts N ROI features $F_{key} = \{f_{p1}, f_{p2}, \dots, f_{pN}\}$ for each part on shared feature maps with deviation information. We encode deviation information with a two-channel feature map, where the channels denote the *x* and *y* offsets of each pixel on the feature map to the center point of the object bounding box. Therefore, we then connect it with the image feature map. We utilize two connected layers to parse the scene spatial configuration map to get spatial attention scores $B_{hp} = \{b_{hpi}\} \in \mathbb{R}^N$. ReLU layer is adopted after the first layer, and a Sigmoid layer is used after the second layer to normalize the final prediction to [0,1]. The scores are utilized for weighting the pose-part area features, and the output feature f_b of this branch is got via the concatenation process and two layers of full connection, as shown in equation (5):

$$f_b = FCs((b_{hp1} \otimes f_{p1}) \oplus \dots \oplus (b_{hpN} \otimes f_{pN})) \tag{5}$$

where $\{b_{hpi} \in [0,1]\}_{i=1}^{N}$, ⊕ indicates the concatenation process, ⊗ indicates element-wise multiplication, $\{f_{pi}\}_{i=1}^{N}$ mean the pose-parts features with deviation information and FCs represents two fully connected layers.

C. Spatial Graph Branch

To generate effective features containing visual and spatial information for human-object pairs, in this branch, we propose to use humans and objects as nodes and their relationships as edges to construct a graph structure and utilize graph neural networks for parsing.

We construct a bipartite graph $\mathcal{G} = (\mathcal{H}, \mathcal{O}, \mathcal{E})$ instead of a fully connected graph. This simplification is to avoid unnecessary calculations. One side of the graph is the p_i set H with $c_i = h$, and the other is the p_i set O with $c_i \neq h$, where *h* indicates the object is human or not. Edge set $\mathcal{E}$ connects all the detected humans and objects, as shown in Fig. 5. Instance-level visual appearance features and spatial information features determine the weight of the edge between the interaction pairs. The parsing module generates the feature message for propagation in the graph neural network according to the sender's visual appearance features and the encoded spatial information features. To construct the graph, we use Faster RCNN [11] as an object detector and apply appropriate filtering to obtain the detection candidate set $\{p_i = (b_i, s_i, c_i)\}_{i=1}^{n}$. b_i denotes the bounding box with dimension 4, si P r0, 1s represents the bounding box with dimension 4, $s_i \in [0, 1]$ represents the score given by the detector, and $c_i \in C$ represents the object category of the candidate. We utilize RoIAlign [35] to extract the visual features of the candidate set as node features, and the calculated spatial feature vectors are used as edge features.

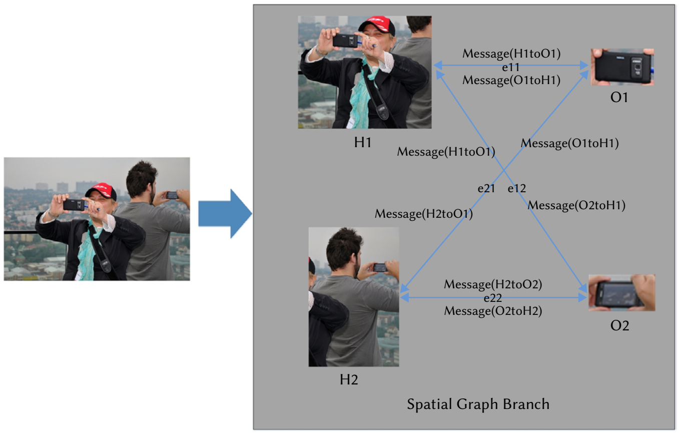

Fig. 5. Spatial graph branch. This branch takes humans and objects as two kinds of nodes to form a bipartite graph structure.

We calculate the human-object information in the image space as the spatial feature, including the center coordinates, width, height, aspect ratio, intersection and area of human and object bounding boxes, and normalize by the image size. d_x is the difference between the center point coordinates on the horizontal axis. d_y is the difference of the center point coordinates on the vertical axis. After normalization by the size of human bounding box, we utilize [$ReLU(d_x)$, $ReLU(\text{-}d_x)$, $ReLU(d_y)$, $ReLU(\text{-}d_y)$] to express the orientation of the bounding box pair. Then, we obtain the spatial information feature vector f_{sp} with dimension 18. Regarding the work of Gupta et al. [38], we connect it with its logarithm to learn higher-order combinations of different terms as equation (6):

$$f'_{sp} = f_{sp} \oplus log(f_{sp} + \theta) \tag{6}$$

where θ is a small constant greater than 0, $\oplus$ denotes concatenation operation, and log represents a logarithmic operation. Then, we utilize a fully connected layer to transform f'_{sp} to the same dimensional space as f_h and f_o.

We adopt the graph neural network and propagate the feature message determined by the sender and spatial information features. Therefore, even if it is the same sender, the feature message will differ depending on the receiver. The message parsing module is defined as equations (7) and (8):

$$M_{oh}(f_{oj}, e_{ij}) = FCs(f_{oj} \oplus e_{ij}) \tag{7}$$

$$M_{ho}(f_{hi}, e_{ij}) = FCs(f_{hi} \oplus e_{ij}) \tag{8}$$

where $\oplus$ represents the concatenation process, and FCs represents two layers of full connection. The weights of the human node as the sender and the object node as the sender are not shared. f_{oj} is the node feature of the j-th object ($j \in \{1, \ldots, |\mathcal{O}|\}$), f_hi the node feature of the i-th human ($i \in \{1, \ldots, |\mathcal{H}|\}$), e_{ij} is f'_{sp} of the pair of the *i*-th human and the *j*-th object.

According to the given feature message and edge weights, the process of transferring feature messages in the graph structure is defined as equations (9) and (10):

$$f'_{hi} = f_{hi} + \sum_{j=1}^{\mathcal{O}} \alpha_{ij} \, M_{oh}(f_{oj}, e_{ij}) \tag{9}$$

$$f'_{oj} = f_{oj} + \sum_{i=1}^{\mathcal{H}} \alpha_{ji} \, M_{ho}(f_{hi}, e_{ij}) \tag{10}$$

where α_{ij} and α_{ji} denote the weight relationship between h_i and o_j. Different from previous works [39], [40], the weight of edges is defined as the visual similarity. In our proposed method, the weight of edges is determined by instance-level visual features and spatial information features. Therefore, the edge weight value of the pair of the *i*-th human and the *j*-th object is calculated as equation (11):

$$\alpha_k = FCs(f_{hi} \oplus f_{oj} \oplus e_{ij}) \tag{11}$$

where $\oplus$ represents concatenation operation, and *FCs* represents two fully connected layers, $k \in \{1, \ldots, |\mathcal{O} \times \mathcal{H}|\}$. A ReLU layer and Sigmoid function are used following the first and second layers. The edge weight value α_{ij} is obtained by applying softmax according to the number of connected object nodes and α_k. Similarly, α_{ji} is normalized according to the number of connected human nodes and α_k. Finally, we pair the graph features and calculate the output features f_c as equation (12):

$$f_c = FCs(f'_{hi} \oplus f'_{oj}) \tag{12}$$

where $\oplus$ represents the concatenation process, and *FCs* represents two layers of full connection.

D. Training and Inference

We obtain the triples of ⟨*human* (*h*), *interaction* (*i*), *object* (*o*)⟩ to calculate the HOI score $S^i_{h,o}$ as the final output of our model. We evaluate triples from two aspects: first, use instance-level visual features and context features to quantify whether the human-object has a relationship with the value of the relationship score s_r; second, a human usually executes multiple interactions as a multi-label classification problem. So we aggregate extra information from scene-level and part-level visual branches, and apply the Sigmoid function to each interaction category to obtain scores $S_{ps} = \{s^i_{ps}\} \left(\forall i \in \mathcal{I} = \{1, \ldots, M\}\right)$, where $\mathcal{I}$ is a collection of interaction categories, and M denotes the total of interaction categories. The two-stage fusion strategy utilizes s_r to inhibit considerable background pairings and enhance the detection precision. They are calculated as equations (13) and (14):

$$s_r = \sigma(FCs(f_a)) \tag{13}$$

$$S_{ps} = \sigma(FCs(f_a \oplus f_b \oplus f_c)) \tag{14}$$

Here, $\oplus$ represents the concatenation process, and σ is a sigmoid function that follows the two layers of full connection. The result features of the three branches are denoted by f_a, f_b, and f_c, respectively.

We utilize the humans and objects with high scores obtained in the object detection stage to form an initial bipartite graph during the training process. Therefore, we combined the two evaluation indicators described above with the detection scores of the human-object pair to obtain the HOI scores $S^i_{h,o}$, as shown in equation (15):

$$S^i_{h,o} = s_r \cdot s^i_{ps} \cdot s_h \cdot s_o. \tag{15}$$

Here, s_h and s_o are the scores of humans and objects obtained in the object detection stage.

Because the HOI tags in the different datasets are unbalanced, we utilize the weighted binary cross-entropy loss $\mathcal{L}_{cls} = w_p \cdot y \cdot log(\hat{y}) + w_n \cdot (1 - y) \cdot log(1 - \hat{y})$, where w_p and w_n denote the weight ratios for positively and negatively samples. Our loss function expression is as equation (16):

$$\mathcal{L} = \sum_{i=1}^{C} \mathcal{L}_{cls}\,(t^i, s^i_{ps}) + \lambda \mathcal{L}_{cls}(r, s_r) \tag{16}$$

where λ represents the weight used to regulate the impact of the loss term, $T = \{t^i \in [0{,}1]\}$ denotes the interaction categories label collection, and r represents the interaction relationship label collection. Moreover, t^i denotes the ground truth relationship label of the i-th interaction of the sample, and $r \in \{0{,}1\}$ represents the presence of the interaction of the pairing.

IV. Experiments

In this part, we describe our experimental result. We begin by explaining the datasets and evaluation metrics along with our implementation details. Then, we perform an extensive quantitative analysis of our proposed model to prove the effectiveness of our approach. Eventually, we utilize ablation experiments to illustrate the influence of these branches in our approach.

A. Datasets and Metric

Datasets. To evaluate the performance of our model, we use two benchmarks for HOI detection: V-COCO [5] and HICO-DET [13]. V-COCO has derived from MS-COCO [41] dataset. It has 10,346 images and 16,199 human instances (2533 images are contained in the training set, 2867 images are for validating, and 4,946 in the test set). The V-COCO dataset contains 26 binary interaction categories. If the object in the image is related to the action, the object is also be annotated. HICO-DET contains 38,118 training images and 9,658 test images with bounding box annotations, 600 HOI categories for 80 object classes (the same as those in the MS COCO data set [41]) and 117 action verbs.

Evaluation metric. For these two data sets, we adopt the evaluation settings in [5]. The results are reported in terms of role mean average precision (*mAProle*). In the research, the purpose is to detect the triples of ⟨*human, interaction, object*⟩. A detected triplet is deemed as a true positive if it has the correct action label, and the minimum of human overlap IOU_h and object overlap IOU_o is greater than 0.5. To demonstrate the effectiveness of our proposed method in interactions with different numbers of annotations, we follow previous practices [13], and the report is divided into three different HOI category sets for the HICO-DET dataset: (a) all 600 HOI categories in HICO (Full), (b) 138 HOI categories with less than ten training instances (Rare), and (c) 462 HOI categories with ten or more training instances (Non-Rare).

B. Implementation Details

As previously stated, we use Faster RCNN [11] and ResNet-50-FPN [42] backbone to obtain the bounding box prediction of humans and objects and CPN [12] to estimate human pose. These have been pre-trained using the MS-COCO dataset. The pose structure comprises N = 17 key-points, which correspond to the MS-COCO data set [41]. The ROI feature with the highest resolution is obtained from the feature map in FPN [42]. In the object detection stage, first, we remove the candidate boxes with a score lower than 0.2 and perform a non-maximum suppression (NMS) operation with a parameter of 0.5. After that, we sort the candidate boxes and select the top 15 humans and objects, respectively, to form a bipartite graph and remove the candidate pairs that contain the same human twice. The resolution of the RoIAlign algorithm in the instance level visual branch is $R_h = 7$, through a residual block, and then global average pooling is similar to [7]. After these steps, we obtain three feature vectors of human, object and context with size $R = 256$, size $D = 3R$. In the part level visual branch, the RoIAlign algorithm produces $5 \times 5 \times (R + 2)$ output features for every region and then utilizes a residual block and GAP to downsize it to $1 \times 1 \times (R + 2$ size. To train the model, we use SGD as the optimizer, with a momentum of 0.9 and weight decay of 1e-4. All data sets have an initial learning rate of 4e-2. Our model has trained 36k iterations and 250k iterations on V-COCO and HICO-DET, respectively. Furthermore, for these two data sets, we reduced the learning rate to 4e-3 at iteration 18k and iteration 200k, respectively.

C. Results

Quantitative results on the two test datasets and the performance comparison between our approach and the current approaches are shown in Tables I and II. We set the baseline to include only the four streams in the instance level visual branch. Our final method integrates all the branches and components introduced in section III.

From Table I we can see that we compare our model with the current ten approaches [5]–[8], [40], [43]–[47] on V-COCO dataset. In the existing work, the GPNN method [40] utilizes the graph neural network to learn and detect interactions, reaching a mAP_{role} of 44.0. The iCAN model [7] integrates three visual feature streams using the attention mechanism in the early fusion approach and provides a mAP_{role} of 45.3. RPNN [44] utilizes a wide range of part-level visual features to detect interactions and realizes a mAP_{role} of 47.53. Zhou et al. [46] introduces a cascade architecture for HOI understanding from coarse to fine and achieves a mAP_{role} of 48.9. PMFNet [47] further divides the part-level visual feature into the same number of pose keypoints and obtains a mAP_{role} of 52.0. Our baseline method achieves 49.3 mAP, and the full method performs the highest effect of 52.8 mAP. As shown in Table II, following the evaluation metrics provided in [13], our model is evaluated on three different HOI categories, namely full, rare, and non-rare with default settings. The full proposed method acquires a steady 20.31 mAP on the HICO-DET test dataset, which could attribute to representative body-part features and potential relationship features in the scene.

D. Qualitative Results

Fig. 6 shows the qualitative outcomes and compares the HOIs detection results of our final (blue scores) and baseline (yellow scores) models. The representative human-object pairs in these images contain variance in object size, human body sizes, and different interaction categories. The interaction prediction probabilities of the correct interaction are visualized. For difficult HOIs, we observe that our final approach yields more dependable results.

Even when the crowd is dense and the spatial distribution of the crowd is uniform, or, the interaction is subtle and the object is tiny and interacts with some representative body part, SMPNet improves the score based on the baseline performs well. This shows that additional information is provided for these categories of interaction from visual features at the part and scene levels.

Special cases: Fig. 7 illustrates our proposed method's success and failure cases in multi-person and multi-object scenes. In Fig. 7(a) and Fig. 7(b), the interactive objects are the same categories. Due to the suppression of representative key parts and scene spatial information, in these two figures, (1,2) and (3,4) human-object pairs have obtained high values, and the values of the (1,4) and (3,2) human-object pairs are all approaching 0. However, when the visual or spatial overlap is high, and the representative parts and the spatial information of the scene are both confusing, the proposed method also products to erroneous predictions. Fig. 7(c) corresponds to the (3,6),(5,8), (5,4), and (7,6) human-object pairs in Table III, and Fig. 7(d) corresponds to the (1,4) and (2,3) human-object pairs in Table IV have obtained high values. At the same time, due to the suppression of representative key parts and scene spatial information, the values of (5, 4) and (3, 6) human-object pairs in Table IV are suppressed.

TABLE I. Performance Comparison on the V-COCO [5] Dataset. The Most Competitive Methods in Each Category Dataset Are Shown in Bold

Method	Feature Backbone	mAP_{role}
Gupta et al. [5]	ResNet-50-FPN	31.8
InteractNet [6]	ResNet-50-FPN	40.0
Kolesnikov et al. [43]	ResNet-50	41.0
GPNN [40]	Deformable ConvNets [48]	44.0
iCAN [7]	ResNet-50	45.3
Wang et al. [8]	ResNet-50	47.3
RPNN [44]	ResNet-50	47.5
Li et al. [45]	ResNet-50-FPN	47.8
Zhou et al. [46]	ResNet-50	48.9
PMFNet [47]	ResNet-50-FPN	52.0
Our baseline	ResNet-50-FPN	49.3
Our method	ResNet-50-FPN	**52.8**

TABLE II. The HOI Detection Performance on the HICO-DET [13] Test Set With the Default Setting (MAP×100). The Most Competitive Methods in Each Category Dataset Are Shown in Bold

		Default		
Method	Feature Backbone	Full	Rare	Non-rare
Shen et al. [49]	VGG-19	6.46	4.24	7.12
InteractNet [6]	ResNet-50-FPN	9.94	7.16	10.77
GPNN [40]	Deformable ConvNets [48]	13.11	9.34	14.23
iCAN [7]	ResNet-50	14.84	10.45	16.15
Wang et al. [8]	ResNet-50	16.24	11.16	17.75
RPNN [44]	ResNet-50	17.35	12.78	18.71
PMFNet [47]	ResNet-50-FPN	17.46	15.65	18.00
Our baseline	ResNet-50-FPN	15.68	12.82	16.54
Our method	ResNet-50-FPN	**20.31**	**17.14**	**21.26**

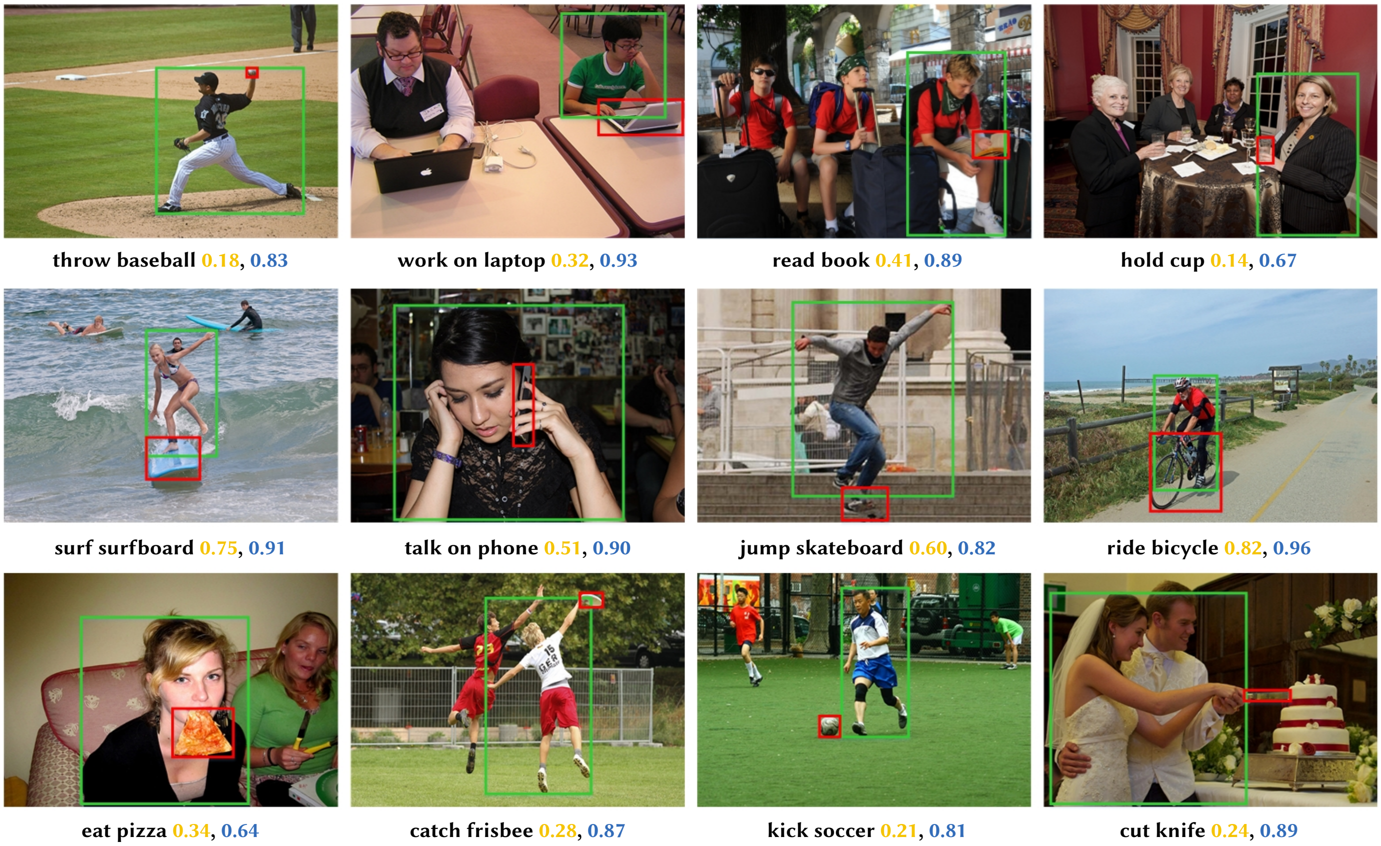
throw baseball 0.18, 0.83
work on laptop 0.32, 0.93
read book 0.41, 0.89
hold cup 0.14, 0.67
surf surfboard 0.75, 0.91
talk on phone 0.51, 0.90
jump skateboard 0.60, 0.82
ride bicycle 0.82, 0.96
eat pizza 0.34, 0.64
catch frisbee 0.28, 0.87
kick soccer 0.21, 0.81
cut knife 0.24, 0.89

Fig. 6. The qualitative output of the final and baseline approaches on the V-COCO [5] test set. Yellow values and blue values denote scores predicted by the base model (instance level visual branch only) and SMPNet, respectively.

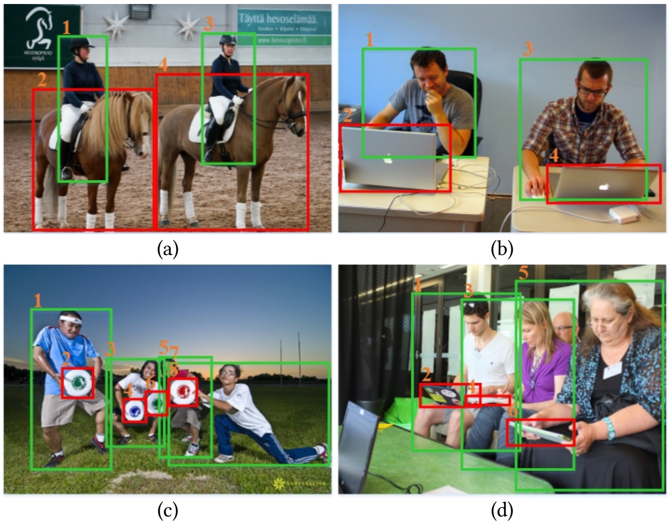

Fig. 7. The success(a and b) and failure(c and d) results of our proposed method for HOI detection. The scores of human-object pairs in failure results (c) and (d) correspond to Tables III and IV, respectively.

TABLE III. The “Hold Frisbee” Interaction Score of Human-Object Pairs in Fig. 7(C)

Instance number	1	3	5	7
2	0.8542	0.0017	0.0009	0.0002
4	0.0023	0.5621	0.3821	0.0042
6	0.0008	0.5308	0.5025	0.3831
8	0.0006	0.0026	0.4499	0.7259

TABLE IV. The “Look” Interaction Score of Human-Object Pairs in Fig. 7(D)

Instance number	1	3	5
2	0.6725	0.3899	0.0008
4	0.4335	0.8214	0.0026
6	0.0015	0.0056	0.8572

E. Ablation Studies

In this section, we conduct ablation study experiments using the V-COCO dataset to assess the efficiency of the different components of the proposed model. As formerly mentioned, we consider the basic model as an instance-level visual branch without part and scene levels, as in [7].

Spatial attention scores. We refer to PLVB when a variant of the part-level visual branch is analyzed without the spatial attention scores from the spatial configuration map. This branch does not use refined pose-part area features but directly utilizes the features obtained by the ROIAlign algorithm on shared feature maps for concatenation operation. We call the branch network using spatial attention scores as S-PLVB. As shown in Table V, spatial attention scores enhance HOI detection ability by 0.4 mAP.

TABLE V. Results of Ablation Studies on the V-COCO Dataset

Model	mAP_{role}
Baseline	49.3
Baseline+PLVB	50.7
Baseline+S-PLVB	51.1
Baseline+SGB	50.9
Baseline+S-PLVB+SGB	52.5
Our method (Baseline+S-PLVB+SGB+RS)	52.8

Part-level visual branch with spatial attention (S-PLVB). This is the vital component. Enlarging and capturing the features of the human pose-part area can effectively obtain relevant information of representative critical parts in interactions. A variation of our model is constructed to assess the impact of this branch. In comparison with the results of the basic model, the mAP utilizing S-PLVB is significantly improved from 49.3 mAP to 51.1 mAP, as shown in Table V.

Spatial graph branch (SGB). In this model branch, we construct a bipartite graph simulating scene with instances as nodes and obtain multiple human-object pairs relationship features in the whole scene through it. A variant of the model we proposed is executed without this branch. Compared with other results, the experiment shows an improvement of 1.6 mAP, as shown in Table V.

Relationship score (RS). Similarly to [47], we utilize the s_r to estimate the existence of an interactive relationship between humans and objects. Its purpose is to inhibit the score value without an interactive relationship. We state that the relationship score is based on the entire existence of the S-PLVB and SGB. In other situations, we straight fuse the results features of the three branches. The RS enhances performance by 0.3 mAP, as shown in Table V.

V. Conclusion

In our research, we proposed an effective human-object interaction detection model SMPNet, that utilizes a multi-level feature parsing strategy. It uses the instance, part, and scene levels parsing branches. In addition to the usual instance-level visual features, we introduce the pose of each interaction instance and the features of the keypoints' region, and utilize the spatial configuration map to generate spatial attention features to refine the visual features of these two levels. In the scene-level branch, we use graph neural networks to simulate the interaction between pairs in the entire scene, and add the spatial information of human-object pairs to adjust visual features. Finally, using the V-COCO and HICO-DET datasets, we demonstrate that our proposed model greatly increases detection capability and exceeds state-of-the-art techniques.

Acknowledgment

This work is supported by the National Natural Science Foundation of China (62072094) and the LiaoNing Revitalization Talents Program (XLYC2005001).

References

[1] P. Wang, P. Chen, Y. Yuan, D. Liu, Z. Huang, X. Hou, G. Cottrell, "Understanding convolution for semantic segmentation," in *2018 IEEE winter conference on applications of computer vision (WACV)*, 2018, pp. 1451– 1460, IEEE.

[2] J. Lu, M. Nguyen, W. Q. Yan, "Deep learning methods for human behavior recognition," in *2020 35th International Conference on Image and Vision Computing New Zealand (IVCNZ)*, 2020, pp. 1–6, IEEE.

[3] L. Mi, Z. Chen, "Hierarchical graph attention network for visual relationship detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 13886–13895.

[4] A. Gupta, A. Kembhavi, L. S. Davis, "Observing human-object interactions: Using spatial and functional compatibility for recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 31, no. 10, pp. 1775–1789, 2009.

[5] S. Gupta, J. Malik, "Visual semantic role labeling," *arXiv preprint arXiv:1505.04474*, 2015.

[6] G. Gkioxari, R. Girshick, P. Dollár, K. He, "Detecting and recognizing human-object interactions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8359–8367.

[7] C. Gao, Y. Zou, J.-B. Huang, "ican: Instance- centric attention network for human-object interaction detection," *arXiv preprint arXiv:1808.10437*, 2018.

[8] T. Wang, R. M. Anwer, M. H. Khan, F. S. Khan, Y. Pang, L. Shao, J. Laaksonen, "Deep contextual attention for human-object interaction detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 5694–5702.

[9] A. Bansal, S. S. Rambhatla, A. Shrivastava, R. Chellappa, "Spatial priming for detecting human-object interactions," *arXiv preprint arXiv:2004.04851*, 2020.

[10] Y. Liao, S. Liu, F. Wang, Y. Chen, C. Qian, J. 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*, 2020, pp. 482–490.

[11] S. Ren, K. He, R. Girshick, J. Sun, "Faster r-cnn: towards real-time object detection with region proposal networks," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 6, pp. 1137–1149, 2016.

[12] Y. Chen, Z. Wang, Y. Peng, Z. Zhang, G. Yu, J. Sun, "Cascaded pyramid network for multi-person pose estimation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7103– 7112.

[13] Y.-W. Chao, Y. Liu, X. Liu, H. Zeng, J. Deng, "Learning to detect human-object interactions," in *2018 ieee winter conference on applications of computer vision (wacv)*, 2018, pp. 381–389, IEEE.

[14] V. Mnih, N. Heess, A. Graves, *et al.*, "Recurrent models of visual attention," in *Advances in neural information processing systems*, 2014, pp. 2204–2212.

[15] R. Yu, K. Yang, B. Guo, "The interaction graph auto-encoder network based on topology-aware for transferable recommendation," in *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, 2022, pp. 2403–2412.

[16] R. Yu, B. Guo, K. Yang, "Selective prototype network for few-shot metal surface defect segmentation," *IEEE Transactions on Instrumentation and Measurement*, vol. 71, pp. 1–10, 2022.

[17] B. Guo, Y. Wang, S. Zhen, R. Yu, Z. Su, "Speed: Semantic prior and extremely efficient dilated convolution network for real-time metal surface defects detection," *IEEE Transactions on Industrial Informatics*, vol. 19, no. 12, pp. 11380-11390, 2023.

[18] D. Bahdanau, K. Cho, Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[19] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, "Advances in neural information processing systems," *Proceedings of Machine Learning Research*, pp. 5998–6008, 2017.

[20] N. Dalal, B. Triggs, "Histograms of oriented gradients for human detection," in *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, vol. 1, 2005, pp. 886–893, IEEE.

[21] S. R. Sain, "The nature of statistical learning theory," *Technometrics*, vol. 38, no. 4, pp. 409, 1996.

[22] Y. Freund, R. E. Schapire, *et al.*, "Experiments with a new boosting algorithm," in *icml*, vol. 96, 1996, pp. 148– 156, Citeseer.

[23] R. Girshick, J. Donahue, T. Darrell, J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *CVPR*, 2014.

[24] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1440–1448.

[25] J. Redmon, S. Divvala, R. Girshick, A. Farhadi, "You only look once: Unified, real-time object detection," 2015, https://doi.org/10.48550/arXiv.1506.02640.

[26] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Y. Fu, A. C. Berg, "SSD: Single shot multibox detector," *Lecture Notes in Computer Science*, vol 9905, pp 21–37, 2016.

[27] J. Johnson, R. Krishna, M. Stark, L.-J. Li, D. Shamma, M. Bernstein, L. Fei-Fei, "Image retrieval using scene graphs," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3668–3678.

[28] J. Yang, J. Lu, S. Lee, D. Batra, D. Parikh, "Graph r- cnn for scene graph generation," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 670–685.

[29] H. Zhang, Z. Kyaw, S.-F. Chang, T.-S. Chua, "Visual translation embedding network for visual relation detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 5532– 5540.

[30] B. Xu, Y. Wong, J. Li, Q. Zhao, M. S. Kankanhalli, "Learning to detect human-object interactions with knowledge," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2019-2028.

[31] S. Wang, K.-H. Yap, J. Yuan, Y.-P. Tan, "Discovering human interactions with novel objects via zero-shot learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 11652–11661.

[32] A. Bansal, S. S. Rambhatla, A. Shrivastava, R. Chellappa, "Detecting human-object interactions via functional generalization," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, 2020, pp. 10460–10469.

[33] S. Gao, H. Wang, J. Song, F. Xu, F. Zou, "An improved human-object interaction detection network," in *2019 IEEE 13th International Conference on Anti-counterfeiting, Security, and Identification (ASID)*, 2019, pp. 192–196, IEEE.

[34] K. He, X. Zhang, S. Ren, J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[35] K. He, G. Gkioxari, P. Dollár, R. Girshick, "Mask r- cnn," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961–2969.

[36] Z. Su, Y. Wang, Q. Xie, R. Yu, "Pose graph parsing network for human-object interaction detection," *Neurocomputing*, vol. 476, pp. 53-62, 2022.

[37] Y.-L. Li, S. Zhou, X. Huang, L. Xu, Z. Ma, H.-S. Fang, Y. Wang, C. Lu, "Transferable interactiveness knowledge for human-object interaction detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3585–3594.

[38] T. Gupta, A. Schwing, D. Hoiem, "No-frills human- object interaction detection: Factorization, layout encodings, and training techniques," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 9677–9685.

[39] L. Li, Z. Gan, Y. Cheng, J. Liu, "Relation-aware graph attention network for visual question answering," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 10313–10322.

[40] S. Qi, W. Wang, B. Jia, J. Shen, S.-C. Zhu, "Learning human-object interactions by graph parsing neural networks," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 401–417.

[41] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C. L. Zitnick, "Microsoft coco: Common objects in context," in *European conference on computer vision*, 2014, pp. 740–755, Springer.

[42] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2117– 2125.

[43] A. Kolesnikov, A. Kuznetsova, C. Lampert, V. Ferrari, "Detecting visual relationships using box attention," in *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 2019, pp. 1749-1753.

[44] P. Zhou, M. Chi, "Relation parsing neural network for human-object interaction detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 843–851.

[45] Y.-L. Li, S. Zhou, X. Huang, L. Xu, Z. Ma, H.-S. Fang, Y. Wang, C. Lu, "Transferable interactiveness knowledge for human-object interaction detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3585–3594.

[46] T. Zhou, W. Wang, S. Qi, H. Ling, J. Shen, "Cascaded human-object interaction recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 4263–4272.

[47] B. Wan, D. Zhou, Y. Liu, R. Li, X. He, "Pose- aware multi-level feature network for human object interaction detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 9469–9478.

[48] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, Y. Wei, "Deformable convolutional networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 764–773.

[49] L. Shen, S. Yeung, J. Hoffman, G. Mori, L. Fei-Fei, "Scaling human-object interaction recognition through zero-shot learning," in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2018, pp. 1568–1576, IEEE.

Zhan Su

Zhan Su received his B.S. and M.S. degree in software engineering from Northeastern University, Shenyang, China, in 2015 and 2017, respectively. He is currently a Ph.D. candidate at Northeastern University, Shenyang, China. His research interests include computer vision, machine learning, and action detection.

Ruiyun Yu

Ruiyun Yu is currently a professor and vice dean of the Software College at the Northeastern University, China. He received his Ph.D. and M.S. degree in computer science and bachelor degree in Mechanical Engineering from the Northeastern University in 2009, 2004, and 1997, respectively. He serves as the director of center for Cross-media Artificial Intelligence. He is one of the Baiqianwan Talents of Liaoning Province, China (Hundred Talents Level), and now a member of the CCF IoT Committee, and a Senior Member of CCF. His research interests include intelligent sensing and computing, computer vision, data intelligence, etc.

Shihao Zou

Shihao Zou received the B.Sc. degree from Beijing Institute of Technology, China, in 2017, and the M.Res. degree from University College London, UK, in 2018. He is currently a Ph.D. candidate at University of Alberta. His interests include computer vision and machine learning, especially human pose and shape estimation, motion capture system.

Bingyang Guo

Bingyang Guo is currently a PhD candidate in the software college of Northeastern University China. He received his Bachelor degree in mechanical engineering from Shenyang Ligong University, China, in 2018, and Master degree in mechanical design and theory from Northeastern University, China, in 2021. His research focuses on image segmentation, image restoration and defect detection.

Li Cheng

Li Cheng received the Ph.D. degree in computer science from the University of Alberta, Canada. He is an associate professor with the Department of Electrical and Computer Engineering, University of Alberta. He has previously worked at A*STAR, Singapore, TTI-Chicago, USA, and NICTA, Australia. His research expertise is mainly on computer vision and machine learning. He is a senior member of the IEEE.