Pose-aware Multi-level Feature Network for Human Object Interaction Detection

Bo Wan* Desen Zhou* Yongfei Liu Rongjie Li Xuming He
ShanghaiTech University, Shanghai, China
{wanbo, zhouds, liuyf3, lirj2, hexm}@shanghaitech.edu.cn

Abstract

Reasoning human object interactions is a core problem in human-centric scene understanding and detecting such relations poses a unique challenge to vision systems due to large variations in human-object configurations, multiple co-occurring relation instances and subtle visual difference between relation categories. To address those challenges, we propose a multi-level relation detection strategy that utilizes human pose cues to capture global spatial configurations of relations and as an attention mechanism to dynamically zoom into relevant regions at human part level. Specifically, we develop a multi-branch deep network to learn a pose-augmented relation representation at three semantic levels, incorporating interaction context, object features and detailed semantic part cues. As a result, our approach is capable of generating robust predictions on fine-grained human object interactions with interpretable outputs. Extensive experimental evaluations on public benchmarks show that our model outperforms prior methods by a considerable margin, demonstrating its efficacy in handling complex scenes. Code is available at https://github.com/bobwan1995/PMFNet.

1. Introduction

Visual relations play an essential role in a deeper understanding of visual scenes, which usually requires reasoning beyond merely recognizing individual scene entities [22, 15, 30]. Among different types of visual relations, human object interactions are ubiquitous in our visual environment and hence its inference is critical for many vision tasks, such as activity analysis [1], video understanding [29] and visual question answering [10].

The task of human object interaction (HOI) detection aims to localize and classify triplets of human, object and relation from an input image. While deep neural networks have led to significant progresses in object and action recognition [13, 21, 6], it remains challenging to detect HOIs due to large variations of human-object appearance and spatial configurations, multiple co-existing relations and subtle differences between similar relations [11, 2].

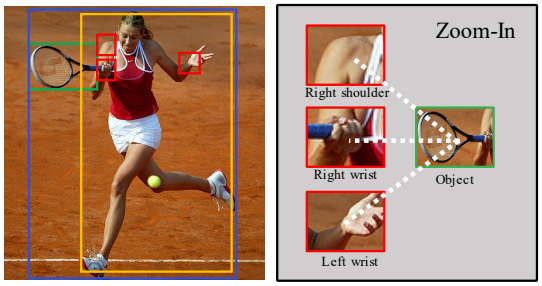

Figure 1. Our framework utilizes three levels of representation, including i): interaction (blue box), ii): visual object (green box and yellow box), iii): human parts (red boxes) to recognize interaction. The highlight of our framework is *human parts* level representation, which can provide discriminative feature. Here several informative human parts, like 'right shoulder', 'right wrist' and 'left wrist', are focused to help recognize action 'hold'.

Most of existing works on HOI detection tackle the problem by reasoning interactions at the visual object level [9, 7, 20]. The dominant approaches typically start from a set of human-object proposals, and extract visual features of human and object instances which are combined with their spatial cues (e.g., masks of proposals) to predict relation classes of those human-object pairs [7, 25, 16]. Despite their encouraging results, such coarse-level reasoning suffers from several drawbacks when handling relatively complex relations. First, it is difficult to determine the relatedness of a human-object pair instance with an object-level representation due to lack of context cues, which can lead to erroneous association. In addition, many relation types are defined in terms of fine-grained actions, which are unlikely to be differentiated based on similar object-level features. For instance, it may require a set of detailed local features to tell the difference between '*hold*' and '*catch*' in sport scenes. Furthermore, as these methods largely rely on holistic features, the reasoning process of relations is a blackbox and hard to interpret.

In this work, we propose a new multi-level relation rea-

*Authors contributed equally and are listed in alphabetical order.

soning strategy to address the aforementioned limitations. Our main idea is to utilize estimated human pose to capture global spatial configuration of relations and as a guidance to extract local features at semantic part level for different HOIs. Such augmented representation enables us to incorporate interaction context, human-object and detailed semantic part cues into relation inference, and hence generate robust and fine-grained predictions with interpretable attentions. To this end, we perform relation reasoning at three distinct semantic levels for each human-object proposal: i) interaction, ii) visual objects, and iii) human parts. Fig. 1 illustrates an example of our relation reasoning.

Specifically, at the *interaction* level of a human-object proposal, we take a union region of the human and object instance, which encodes the context of the relation proposal, to produce an affinity score of the human-object pair. This score indicates how likely there exists a visual relation between the human-object pair and helps us to eliminate background proposals. For the *visual object* level, we adopt a common object-level representation as in [2, 7, 16] but augmented by human pose, to encode human-object appearance and their relative positions. The main focus of our design is a new representation at the *human part* level, in which we use the estimated human pose to describe the detailed spatial and appearance cues of the human-object pair. To achieve this, we exploit the correlation between parts and relations to produce a part-level attention, which enable us to focus on sub-regions that are informative to each relation type. In addition, we compute the part locations relative to the object entity to encode a fine-level spatial configuration. Finally, we integrate the HOI cues from all three levels to predict the category of the human-object proposal.

We develop a multi-branch deep neural network to instantiate our multi-level relation reasoning, which consists of four main modules: a backbone module, a holistic module, a zoom-in module and a fusion module. Given an image, the *backbone module* computes its convolution feature map, and generates human-object proposals and spatial configurations. For each proposal, the *holistic module* integrates human, object and their union features, as well as an encoding of human pose and object location. The *zoom-in module* extracts the human part and object features, and produces a part-level attention from the pose layout to enhance relevant part cues. The *fusion module* combines the holistic and part-level representations to generate final scores for HOI categories. We refer to our model as *Pose-aware Multi-level Feature Network* (PMFNet). Given human-object proposals and pose estimation, our deep network is trained in an end-to-end fashion.

We conduct extensive evaluations on two public benchmarks V-COCO and HICO-DET, and outperform the current state-of-the-art by a sizable margin. To better understand our method, we also provide detailed ablative study of our deep network on the V-COCO dataset.

Our main contributions are three-folds:

- We propose a multi-level relation reasoning for human object interaction detection, in which we utilize human pose to capture global configuration and as an attention to extract detailed local appearance cues.
- We develop a modularized network architecture for HOI prediction, which produces an interpretable output based on relation affinity and part attention.
- Our approach achieves the state-of-the-art performance on both V-COCO and HICO-DET benchmarks.

2. Related Work

Visual Relationship Detection. Visual Relation Detection (VRD) [19, 22, 15, 30] aims at detecting the objects and describing their interactions simultaneously for a given image, which is a critical task for achieving visual scene understanding. Lu *et al.* [19] propose to learn language priors from semantic word embedding to fine-tune visual relationships. Zhang *et al.* [30] design a visual translation network to embed objects into low-dimension relation space for tackling visual relationship detection problem. Besides, Xu *et al.* [26] model the visual relationship detection in structured scene as a graph, and propagate messages between objects. In our task, we focus on human-centric relationship detection, which aims to detect human-object interactions.

Human-Object Interaction Detection. Human-Object interaction (HOI) detection is essential for understanding human behaviors in a complex scene. In recent years, researchers have developed several human object interaction dataset, like V-COCO [11] and HICO-DET [2]. Early studies mainly focus on tackling HOIs recognition by utilizing multi-stream information, including human, object appearance, spatial information and human poses. In HO-RCNN [2], Chao *et al.* propose multi-stream to aggregate human, object and spatial configuration information to resolve HOIs detection tasks. Qi *et al.* [20] propose graph parsing neural network (GPNN) to model the structured scene into a graph and propagate messages between each human and object node and classify all nodes and edges for their possible object classes and actions.

There have been several attempts that use human pose for recognizing fine-grained human related actions [5, 7, 16]. Fang *et al.* [5] exploit pair-wise human parts correlation to help to tackle HOIs detection. Li *et al.* [16] explore interactiveness prior existing in multiple datasets and combine human pose and spatial configuration to form pose configuration map. However, those works only take human pose as a spatial constraint between human parts and object, but do not use it to extract zoom-in feature in each part, which provides more detail information for HOI task.

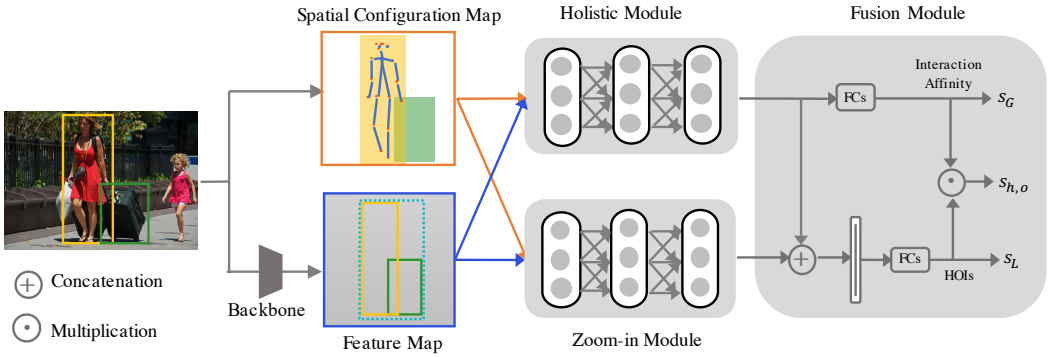

Figure 2. Overview of our framework: For a pair of human-object proposals and related human pose, **Backbone Module** aims to prepare convolution feature map and Spatial Configuration Map (SCM). **Holistic Module** generates object-level features and **Zoom-in Module** captures part-level features. Finally **Fusion Module** combines object-level and part-level cues to predict final scores for HOI categories.

By contrast, we take advantage of this fine-grained feature to capture subtle differences between similar interactions.

Attention Models. Attention mechanism has been proved very effective in various vision tasks, including image captioning [27, 28], fine grained classification [14], pose estimation [4] and action recognition[23, 8]. Attention mechanism can help highlight informative regions or parts and suppress some irrelevant global information. Xu *et al.* [27] firstly utilize attention mechanism in image captioning and automatically attend some informative region in image relevant to generated sentences. Sharma *et al.* [23] apply attention models realized by LSTM into action recognition task to learn important parts of video frames. Yu *et al.* [14] propose a stacked semantic guided attention to focus on informative birds' parts and suppress irrelevant global information. In our work, we focus on pose-aware attention for HOIs detection in human parts.

3. Method

We now introduce our multi-level relation reasoning strategy for human-object interaction detection. Our goal is to localize and recognize human-object interaction instances in an image. To this end, we augment object-level cues with human pose information and propose an expressive relation representation that captures the relation context, human-object and detailed local parts. We develop a multi-branch deep neural network, referred to as PMFNet, to learn such an HOI representation and predict categories of HOI instances. Below we first present an overview of our problem setup and method pipeline in Sec. 3.1, followed by the detailed description of our model architecture in Sec. 3.2. Finally, Sec. 3.3 outlines the model training procedure.

3.1. Overview

Given an image I, the task of human-object interaction detection aims to generate tuples $\{\langle \mathbf{x}_h, \mathbf{x}_o, c_o, a_{h,o} \rangle\}$ for all HOI instances in the image. Here $\mathbf{x}_h \in \mathbb{R}^4$ denotes human instance location (i.e., bounding box parameters), $\mathbf{x}_o \in \mathbb{R}^4$ denotes object instance location, $c_o \in \{1, ..., C\}$ is the object category, and $a_{h,o} \in \{1, ..., A\}$ denotes the interaction class associated with $\mathbf{x}_h$ and $\mathbf{x}_o$. For a pair $\langle \mathbf{x}_h, \mathbf{x}_o \rangle$, we use $c^a_{h,o} \in \{0, 1\}$ to indicate the existence of interaction class a. The object and relation set $\mathcal{C} = \{1, ..., C\}$ and $\mathcal{A} = \{1, ..., A\}$ are given as inputs to the detection task.

We adopt a hypothesize-and-classify strategy in which we first generate a set of human-object proposals and then predict their relation classes. In the proposal generation stage, we apply an object detector (e.g., Faster R-CNN [21]) to the input image and obtain a set of human proposals with detection scores $\{\langle \mathbf{x}_h, s_h \rangle\}$, and object proposals with their categories and detection scores $\{\langle \mathbf{x}_o, c_o, s_o \rangle\}$. Our HOI proposals are generated by pairing up all the human and object proposals. In the relation classification stage, we first estimate a relation score $s^a_{h,o}$ for each interaction a and a given $\langle \mathbf{x}_h, \mathbf{x}_o \rangle$ pair. The relation score is then combined with the detection scores of relation entities (human and object) to produce the final HOI score $R^a_{h,o}$ for the tuple $\{\langle \mathbf{x}_h, \mathbf{x}_o, c_o, a_{h,o} \rangle\}$ as follows,

$$R^a_{h,o} = s^a_{h,o} \cdot s_h \cdot s_o, \tag{1}$$

where we adopt a soft score fusion by incorporating human score s_h and object score s_o at the same time, which represent detection quality of each proposal.

The main focus of this work is to build a pose-aware relation classifier for predicting the relation score $s^a_{h,o}$ given a $\langle \mathbf{x}_h, \mathbf{x}_o \rangle$ pair. To achieve this, we first apply an off-the-shelf pose estimator [3] to a cropped region of proposal $\mathbf{x}_h$, which generates a pose vector $\mathbf{p}_h = \{p^1_h, ..., p^K_h\}$, where $p^k_h \in \mathbb{R}^2$ is k-th joint location and K is the number of all joints. In order to incorporate interaction context, human-object and detailed semantic part cues into relation inference, we then introduce a multi-branch deep neural network to generate

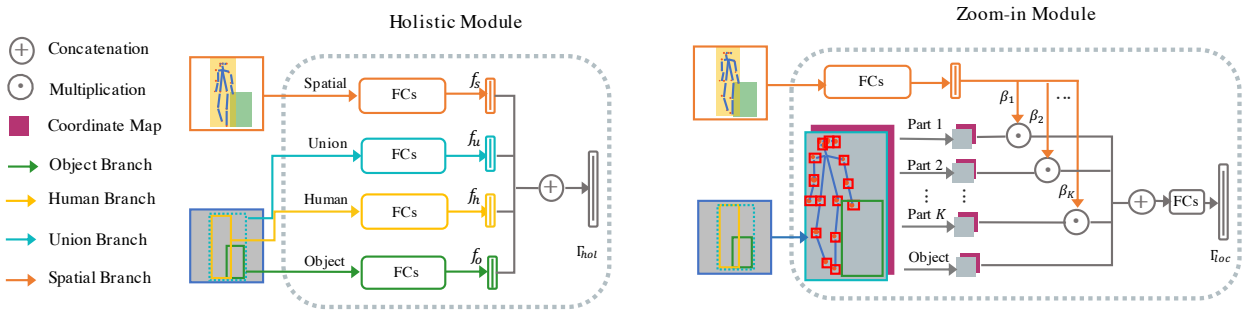

Figure 3. The structure of holistic module and zoom-in module. Holistic module includes human, object, union and spatial branches. Zoom-in module uses human part information and attention mechanism to capture more details.

the relation scores:

$$P(c^a_{h,o} = 1|I) \propto s^a_{h,o} = \mathcal{F}_a(I, \mathbf{x}_h, \mathbf{x}_o, \mathbf{p}_h) \tag{2}$$

where the network $\mathcal{F}_a$ composes of four modules: a backbone module, a holistic module, a zoom-in module and a fusion module. Below we will describe the details of our model architecture.

3.2. Model Architecture

Our deep network, PMFNet, instantiates a multi-level relation reasoning with the following four modules: a) a *backbone* module computes image feature map and generates human-object proposals plus spatial configurations; b) a *holistic* module extracts object-level and context features of the proposals; c) a *zoom-in* module focuses on mining part-level features and interaction patterns between human parts and object; and d) a *fusion* module combines both object-level and part-level features to predict the interaction scores. An overview of our model is shown in Fig. 2.

3.2.1 Backbone Module

We adopt ResNet-50-FPN [17] as our convolutional network to generate feature map $\mathbf{\Gamma}$ with channel dimension of D. For proposal generation, we use Faster R-CNN [21] as object detector to produce relation proposal pairs $\{\langle \mathbf{x}_h, \mathbf{x}_o \rangle\}$. As mentioned earlier, we also compute human pose vector $\mathbf{p}_h$ for each human proposal $\mathbf{x}_h$ and take it as one of the inputs to our network.

In addition to the conv features, we also extract a set of geometric features to encode the spatial configuration of each human-object instance. We start with two binary masks of human and object proposal in their union space to capture object-level spatial configuration as in [2, 7]. Moreover, in order to capture fine-level spatial information of human parts and object, we add an additional pose map with predicted poses following [16]. Specifically, we represent the estimated human pose as a line-graph in which all the joints are connected according to the skeleton configuration of COCO dataset. We rasterize the line-graph using a width of $w = 3$ pixels and a set of intensity values ranging from 0.05 to 0.95 in a uniform interval to indicate different human parts. Finally, the binary masks and pose map in union space are rescaled to $M \times M$ and concatenated in channel-wise to generate a spatial configuration map.

3.2.2 Holistic Module

In order to capture object-level and relation context information, the holistic module is composed of four basic branches: human branch, object branch, union branch and spatial branch, illustrated in Fig. 3 (left). The input features of human, object and union branches are cropped from convolution feature map $\mathbf{\Gamma}$ by applying RoI-Align [12] according to human proposal $\mathbf{x}_h$, object proposal $\mathbf{x}_o$ and their union proposal $\mathbf{x}_u$. $\mathbf{x}_u$ is defined as the minimum box in spatial region that contains both $\mathbf{x}_h$ and $\mathbf{x}_o$. Then human features, object features and union features are rescaled to $R_h \times R_h$ resolution. The input of spatial branch directly comes from spatial configuration map generated in Sec. 3.2.1. For each branch, two fully connected layers are adopted to embed the features to an output feature representation. We denote the output features of human, object, union and spatial feature as f_h, f_o, f_u, f_s, and all the features are concatenated to obtain the final holistic feature Γ_{hol}:

$$\Gamma_{hol} = f_h \oplus f_o \oplus f_u \oplus f_s \tag{3}$$

where $\oplus$ denotes concatenation operation.

3.2.3 Zoom-in Module

While the holistic features provide coarse-level information for interactions, many interaction types are defined at a fine-grained level which require detailed local information of human part or object. Hence we design a zoom-in (ZI) module to zoom into human parts to extract part-level features. The overall zoom-in module can be viewed as a network that takes human pose, object proposal and convolution feature map as inputs and extract a set of local interaction features for the HOI relations:

$$\Gamma_{loc} = \mathcal{F}_{\text{ZI}}(\mathbf{p}_h, \mathbf{x}_o, \mathbf{\Gamma}) \tag{4}$$

Our zoom-in module, illustrated in Fig. 3 (right), consists of three components: i) A part-crop component that aims to extract fine-grained human parts features; ii) A spatial align component that assigns spatial information to human parts features; iii) A semantic attention component that enhances the human part features relevant to interaction and suppress irrelevant ones.

Part-crop component Given the human pose vector $\mathbf{p}_h = \{p_h^1, ..., p_h^K\}$, we define a local region $\mathbf{x}_{p_k} \in \mathbb{R}^4$ for each joint p_h^k, which is a box centered at p_h^k and has a size γ proportional to the size of human proposal $\mathbf{x}_h$. Similar to Sec. 3.2.2, we adopt RoI-Align [12] for those created part boxes together with object proposal $\mathbf{x}_o$ to generate $(K+1)$ regions and rescale to a resolution of $R_p \times R_p$. We denote the pooled part features and object feature as $f_p = \{f_{p_1}, ..., f_{p_K}\}$ and f_{p_o} where each feature is of size $R_p \times R_p \times D$.

Spatial align component Our zoom-in module aims to extract fine-level features of local part regions and model the interaction patterns between human parts and objects. Many interactions have strong correlations with spatial configuration of human parts and object, which can be encoded by the relative locations between different human part and target object. For example, if the target object is close to 'hand', the interaction are more likely to be 'hold' or 'carry', and less likely to be 'kick' or 'jump'. Based on this observation, we introduce the spatial offset of x, y coordinates relative to object center as an additional spatial feature for each part.

In particular, we generate a coordinate map $\boldsymbol{\alpha}$ with the same spatial size as the convolution feature map $\boldsymbol{\Gamma}$. The map $\boldsymbol{\alpha}$ consists of two channels, indicating the x and y coordinates for each pixel in $\boldsymbol{\Gamma}$, and normalized by the object center. Then we apply RoI-Align [12] for each human part $\mathbf{x}_{p_k}$ and object proposal $\mathbf{x}_o$ on $\boldsymbol{\alpha}$ and get the spatial map α_k for part k and α_o for object. We concatenate the spatial map with the part-crop features so that for a $R_p \times R_p$ cropped part region, we align relative spatial offset to each pixel, which augments part features with a fine-grained spatial cues. The final k-th human part feature and object feature are :

$$f'_{p_k} = f_{p_k} \oplus \alpha_k, \quad f'_{p_o} = f_{p_o} \oplus \alpha_o \tag{5}$$

where $f'_{p_k}, f'_{p_o} \in \mathbb{R}^{R_p \times R_p \times (D+2)}$ and $\oplus$ is the concatenate operation.

Semantic attention component As the pose representation also encodes the semantic class of human parts, which typically have strong correlations with interaction types (e.g., 'eyes' are important for 'read' a book). We thus predict a semantic attention using the same spatial configuration map from Sec. 3.2.1.

Our semantic attention network consists of two fully connected layers. A ReLU layer is adopted after the first layer, and a Sigmoid layer is used after the second layer to normalize the final prediction to $[0, 1]$. We denote our inferred semantic attention as $\boldsymbol{\beta} \in \mathbb{R}^K$. Note that we do not predict semantic attention for the object, and assume the object has always an attention of value 1, which means it is uniformly important across different instances. The semantic attention is used to weight the part features as follows:

$$f''_{p_k} = \beta_k \odot f'_{p_k} \tag{6}$$

where $\beta_k \in [0, 1]$ is the k-th value of $\boldsymbol{\beta}$, $\odot$ indicates element-wise multiplication.

Finally, we concatenate human part features and object feature to obtain the attended part-level features f_{att} and feed it to multiple fully-connected layers ($\mathcal{FC}$) to extract final local feature Γ_{loc}:

$$f_{att} = f''_{p_1} \oplus \dots f''_{p_K} \oplus f'_{p_o} \tag{7}$$

$$\Gamma_{loc} = \mathcal{FC}(f_{att}) \tag{8}$$

3.2.4 Fusion Module

In order to compute the score $s^a_{h,o}$ of pair $\langle \mathbf{x}_h, \mathbf{x}_o \rangle$ for each interaction a, we employ a fusion module to fuse relation reasoning from different levels. Our fusion module aims to achieve the following two different goals. First, it uses the coarse-level features as a context cue to determine whether any relation exists for a human-object proposal. This allows us to suppress many background pairs and improve the detection precision. Concretely, we take the holistic feature Γ_{hol} and feed it into a network branch consisting of a two-layer fully-connected network followed by a sigmoid function σ, which generates an interaction affinity score s_G:

$$s_G = \sigma(\mathcal{FC}(\Gamma_{hol})). \tag{9}$$

Second, the fusion module use the object-level and part-level features to determine the relation score based on the fine-grained representation. Using a similar network branch, we compute a local relation score s_L from all the relation features:

$$s^a_L = \sigma(\mathcal{FC}_a(\Gamma_{loc} \oplus \Gamma_{hol})) \tag{10}$$

where a indicates the relation types.

Finally, we fuse those two scores defined above to obtain the relation score for a human-object proposal $\langle \mathbf{x}_h, \mathbf{x}_o \rangle$:

$$s^a_{h,o} = s^a_L \cdot s_G, \quad \forall a \in \mathcal{A}. \tag{11}$$

3.3. Model Learning

In training stage, we freeze ResNet-50 in our backbone module, and train FPN and other components in Sec. 3.2 in

an end-to-end manner. Note that the object detector (Faster R-CNN [21]) and the pose estimator (CPN [3]) are external modules and thus do not participate in learning process.

Assume we have a training set of size N with relation labels set $Y = \{\mathbf{y}^i\}$ and interaction affinity label set $Z = \{z^i\}$, where $\mathbf{y}^i = (y^{1,i}, ..., y^{A,i}) \in \mathbb{1}^A$ indicates ground truth relation label for i-th sample, and $z^i \in \{0, 1\}$ indicates the relatedness of this sample, $i \in \{1, ..., N\}$. We define $z^i = 1$ if $\exists a \in \mathcal{A}, y^{a,i} = 1$, else $z^i = 0$.

Suppose our predicted local relation scores are $S_L = \{\mathbf{s}_L^i\}$ and affinity scores are $S_G = \{s_G^i\}$ for those samples, where $\mathbf{s}_L^i = (s_L^{1,i}, ..., s_L^{A,i})$ indicates the predicted local scores of all interactions and s_G^i is predicted interaction affinity score for i-th sample. As our classification task is actually a multi-label classification problem, we adopt a binary cross entropy loss for each relation class and interaction affinity. Let $L_{CrossEntropy}(a, b) = a\log(b) + (1 - a)\log(1 - b)$, the overall objective function for our training L is defined as:

$$L = \frac{1}{N}\sum_{i=1}^{N}\Big\{\sum_{a=1}^{A} L_{CrossEntropy}(y^{a,i}, s_L^{a,i}) + \mu L_{CrossEntropy}(z^i, s_G^i)\Big\} \quad (12)$$

where μ is a hyperparameter to balance the relative importance of multi-label interaction prediction and binary interaction affinity prediction.

4. Experiments

In this section, we first describe the experimental setting and implementation details. We then evaluate our models with quantitative comparisons to the state-of-the-art approaches, followed by ablation studies to validate the components in our framework. Finally, we show several qualitative results to demonstrate the efficacy of our method.

4.1. Experimental Setting

Datasets We evaluate our method on two HOI benchmarks: V-COCO [11] and HICO-DET [2]. V-COCO is a subset of MS-COCO [18], including 10,346 images (2,533 for training, 2,867 for validation and 4,946 for test) and 16,199 human instances. Each person is annotated with binary labels for 26 action categories. HICO-DET consists of 47,776 images with more than 150K human-object pairs (38,118 images in training set and 9,658 in test set). It has 600 HOI categories over 80 object categories (as in MS-COCO [18]) and 117 unique verbs.

Evaluation Metric We follow the standard evaluation setting in [2] and use mean average precision to measure the HOI detection performance. We consider an HOI detection as true positive when its predicted bounding boxes of both human and object overlap with the ground-truth bounding boxes with IOUs greater than 0.5, and the HOI class prediction is correct.

Methods	AP_{role}
Gupta *et al.* [11]	31.8
InteractNet [9]	40.0
GPNN [20]	44.0
iCAN w/late(early) [7]	44.7 (45.3)
Li *et al.* ($\text{RP}_\text{D}\text{C}_\text{D}$) [16]	47.8
Our baseline	48.6
Our method (PMFNet)	**52.0**

Table 1. Performance comparison on V-COCO [11] test set.

4.2. Implementation Details

We use Faster R-CNN [21] as object detector and CPN [3] as pose estimator, which are pre-trained on the COCO train2017 split. Each human pose has a total of $K = 17$ keypoints as in COCO dataset.

Our backbone module uses ResNet-50-FPN [17] as feature extractor, and we crop RoI features from the highest resolution feature map in FPN [17]. The size of our spatial configuration map M is set to 64. The RoI-Align in holistic module has a resolution $R_h = 7$, while in zoom-in module, the size of human parts is $\gamma = 0.1$ of human box height and all the features are rescaled to $R_p = 5$.

We freeze ResNet-50 backbone and train the parameters of FPN component. We use SGD optimizer for training with initial learning rate 4e-2, weight decay 1e-4, and momentum 0.9. The ratio of positive and negative samples is 1:3. For V-COCO [11], we reduce the learning rate to 4e-3 at iteration 24k, and stop training at iteration 48k. For HICO-DET [20], we reduce the learning rate to 4e-3 at iteration 250k and stop training at iteration 300k. During testing, we use object proposals from [7] for fair comparison. See Suppl. Material for more details.

4.3. Quantitative Results

We compare our proposed framework with several existing approaches for evaluation. We take only human, object and union branches in holistic module as our baseline, while our final model integrates all the modules in Sec. 3.2.

For **V-COCO** dataset, we evaluate AP_{role} of 24 actions with roles as in [11]. As shown in Tab. 1, our baseline method achieves **48.6** mAP, outperforming all the existing approaches [11, 9, 20, 7, 16]. Compared to those methods, our baseline adds a union region feature to capture context information, which turns out to be very effective for predicting interaction patterns in a small dataset like V-COCO. Moreover, our overall model achieves **52.0** mAP, which outperforms all the current state-of-the-art methods by a sizable margin, and further improves our baseline by **3.4** mAP.

For **HICO-DET**, we choose six current state-of-the-art methods [16, 24, 2, 9, 20, 7] for comparison. As shown in Tab. 2, our baseline still performs well and surpasses most existing works except [16]. One potential reason is

Methods	Default			Know Object		
	Full	Rare	Non-Rare	Full	Rare	Non-Rare
Shen *et al.* [24]	6.46	4.24	7.12	-	-	-
HO-RCNN [2]	7.81	5.37	8.54	10.41	8.94	10.85
InteractNet [9]	9.94	7.16	10.77	-	-	-
GPNN [20]	13.11	9.34	14.23	-	-	-
iCAN [7]	14.84	10.45	16.15	16.26	11.33	17.73
Li *et al.*-$\text{RP}_\text{D}\text{C}_\text{D}$ [16]	17.03	13.42	**18.11**	19.17	15.51	20.26
Our baseline	14.92	11.42	15.96	18.83	15.30	19.89
Our method (PMFNet)	**17.46**	**15.65**	18.00	**20.34**	**17.47**	**21.20**

Table 2. Results Comparison on HICO-DET [2] test set.

Methods	Default	
	Interactiveness(520)	No-interaction(80)
Our baseline	15.97	8.05
Our method (PMFNet)	18.79	8.83

Table 3. Improvements of our model in Interactiveness and No-interaction HOIs on HICO-DET [2] test set.

that HICO-DET dataset has a more fine-grained labeling of interactions (117 categories) than V-COCO (24 categories), and hence the object-level cue is insufficient to distinguish the subtle difference between similar interactions. In contrast, our full model achieves the-state-of-art performance with **17.46** mAP and **20.34** mAP on **Default** and **Know Object** categories respectively, outperforming all the existing works. In addition, it further improves our baseline by **2.54** mAP and **1.51** mAP on **Default** and **Know Object** modes respectively.

Furthermore, we divide the 600 HOI categories of the HICO-DET benchmark into two groups as in [16]: Interactiveness (520 non-trivial HOI classes) and No-interaction (80 no-interaction classes for human and each of 80 object categories). We show the performance of our full model on those two groups compared with our baseline in Tab. 3. It is evident that our method achieves larger improvement on the Interactiveness group. As the No-interaction group consists of background classes only, this indicates that our pose-aware dynamic attention is more effective on the challenging task of fine-grained interaction classification.

4.4. Ablation Study

In this section, we perform several experiments to evaluate the effectiveness of our model components on V-COCO dataset (Tab. 4).

Spatial Configuration Map (SCM) As in [16], we augment human and object binary masks with an additional human pose configuration map, which provides more detailed spatial information on human parts. This enriched SCM enables the network to filter out non-interactive human-object instances more effectively. As shown in Tab. 4, the SCM improves our baseline by **0.7** mAP.

Part-crop (PC) Part-crop component zooms into semantic human parts and provides fine-grained feature representation of human body. Experiments in Tab. 4 show the effectiveness of zoom-in parts feature, which improves mAP

Methods	Components					
	SCM	PC	SpAlign	SeAtten	IA	AP_{role}
Baseline	-	-	-	-	-	49.2
Incremental	✓	-	-	-	-	49.9
	✓	✓	-	-	-	51.0
	✓	✓	✓	-	-	52.4
	✓	✓	✓	✓	-	52.7
Drop-one-out	-	✓	✓	✓	✓	52.0
	✓	-	-	-	✓	50.3
	✓	✓	-	✓	✓	51.1
	✓	✓	✓	-	✓	52.6
	✓	✓	✓	✓	-	52.7
Our method (PMFNet)	✓	✓	✓	✓	✓	**53.0**

Table 4. Ablation study on V-COCO [11] val set.

from **49.9** to **51.0**. We note that the following spatial align and semantic attention component are built on top of the part-crop component.

Spatial Align (SpAlign) Spatial align component computes relative locations of all the parts w.r.t. the object and integrates them into the part features, which captures a 'normalized' local context. We observe a significant improvement from **51.0** to **52.4** in Tab. 4.

Semantic Attention (SeAtten) The semantic attention focuses on informative human parts and suppress other irrelevant ones. Its part-wise attention scores provides an interpretable feature for our predictions. As shown in Tab. 4, SeAtten slightly improves the preformance by **0.3** mAP.

Interaction Affinity (IA) Similar to [16], the interaction affinity indicates whether a human-object pair have interaction, and can reduce false positives by lowering their interaction scores. We can observe from Tab. 4 that IA improves performance by **0.3** mAP.

Drop-one-out Ablation study We further perform a drop-one-out ablation study which all independent components are removed individually, shown in tab. 4. The results demonstrate that each independent component indeed contribute to our final performance.

4.5. Qualitative Visualization Results

Fig. 4 shows our HOIs detection results compared with the baseline approach. We can see that our framework is capable of detecting difficult HOIs where the target objects are very small and generates a more confident score. This suggests that part-level features provide more informative visual cues for difficult human-object interaction pairs.

Fig. 5 visualizes semantic attention on a variety of HOI cases, each of which provides an interpretable outcome for our predictions. The highlighted joint regions indicate that our Semantic Attention (SeAtten) component generates an attention score higher than 0.7 for the related keypoint. In Fig. 5(a), for the same person interacting with various target objects, our SeAtten component is capable of automatically focusing on different human parts that are strongly related to interaction action. As the two images in up-left show,

Figure 4. HOI detection results compared with baseline on V-COCO[11] val set. For each ground-truth interaction, we compare interaction action score with baseline method. Red number and green number denote score predicted by baseline and our approach respectively. As shown in figure, our approach can be more confident for predicting interaction actions when the target objects are very small and ambiguous (all score improvements great than 0.5).

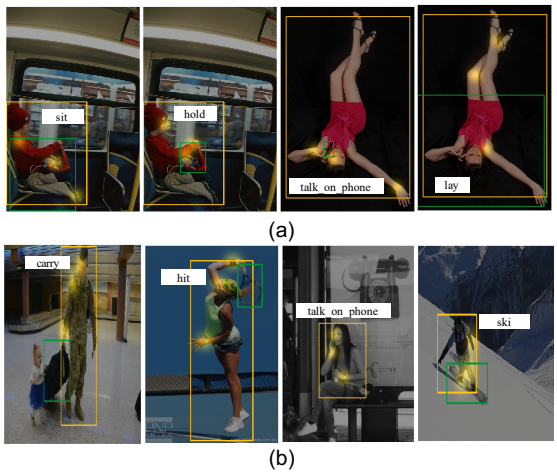

Figure 5. Semantic Attention on (a) the same person interacting with different objects, and (b) different person with various interactions.

when the child interacting with chair, SeAtten will concentrate on the full body joints; while he interacting with an instrument, SeAtten will focus on his hands. To validate the generalization capacity of the SeAtten component, we also visualize several other HOI examples in Fig.5(b). For different persons with various interactions, our SeAtten component can always produce meaningful highlight on human parts which are relevant to each interaction type.

5. Conclusion

In this paper, we have developed an effective multi-level reasoning approach to human-object interaction detection. Our method is capable of incorporating interaction level, visual object level and human parts level features under the guidance of human pose information. As a result, it is able to recognize visual relations with subtle differences. We present a multi-branch deep neural network to instantiate our core idea of multi-level reasoning. Moreover, we introduce a semantic part-based attention mechanism at the part level to automatically extract relevant human parts for each interaction instance. The visualization of our attention map produces an interpretable output for the human-object relation detection task. Finally, we achieve the state-of-the-art performances on both V-COCO and HICO-DET benchmarks, and outperform other approaches by a large margin on V-COCO dataset.

Acknowledgement

This work was supported by Shanghai NSF Grant (No. 18ZR1425100) and NSFC Grant (No. 61703195).

References

[1] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 961–970, 2015.

[2] Yu-Wei Chao, Yunfan Liu, Xieyang Liu, Huayi Zeng, and Jia Deng. Learning to detect human-object interactions. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 381–389. IEEE, 2018.

[3] Yilun Chen, Zhicheng Wang, Yuxiang Peng, Zhiqiang Zhang, Gang Yu, and Jian Sun. Cascaded pyramid network for multi-person pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7103–7112, 2018.

[4] Xiao Chu, Wei Yang, Wanli Ouyang, Cheng Ma, Alan L Yuille, and Xiaogang Wang. Multi-context attention for human pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1831–1840, 2017.

[5] Hao-Shu Fang, Jinkun Cao, Yu-Wing Tai, and Cewu Lu. Pairwise body-part attention for recognizing human-object interactions. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 51–67, 2018.

[6] Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. Convolutional two-stream network fusion for video action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 1933–1941, 2016.

[7] Chen Gao, Yuliang Zou, and Jia-Bin Huang. ican: Instance-centric attention network for human-object interaction detection. In *British Machine Vision Conference (BMVC)*, 2018.

[8] Rohit Girdhar and Deva Ramanan. Attentional pooling for action recognition. In *Advances in Neural Information Processing Systems (NeuIPS)*, pages 34–45, 2017.

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

[10] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

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

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

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 770–778, 2016.

[14] Zhong Ji, Yanwei Fu, Jichang Guo, Yanwei Pang, Zhongfei Mark Zhang, et al. Stacked semantics-guided attention model for fine-grained zero-shot learning. In *Advances in Neural Information Processing Systems (NeuIPS)*, pages 5998–6007, 2018.

[15] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision (IJCV)*, 123(1):32–73, 2017.

[16] Yong-Lu Li, Siyuan Zhou, Xijie Huang, Liang Xu, Ze Ma, Hao-Shu Fang, Yan-Feng Wang, and Cewu Lu. Transferable interactiveness prior for human-object interaction detection. *arXiv preprint arXiv:1811.08264*, 2018.

[17] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2117–2125, 2017.

[18] 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 *European conference on computer vision (ECCV)*, pages 740–755. Springer, 2014.

[19] Cewu Lu, Ranjay Krishna, Michael Bernstein, and Li Fei-Fei. Visual relationship detection with language priors. In *European Conference on Computer Vision (ECCV)*, pages 852–869. Springer, 2016.

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

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

[22] Mohammad Amin Sadeghi and Ali Farhadi. Recognition using visual phrases. In *CVPR 2011*, pages 1745–1752. IEEE, 2011.

[23] Shikhar Sharma, Ryan Kiros, and Ruslan Salakhutdinov. Action recognition using visual attention. *arXiv preprint arXiv:1511.04119*, 2015.

[24] Liyue Shen, Serena Yeung, Judy Hoffman, Greg Mori, and Li Fei-Fei. Scaling human-object interaction recognition through zero-shot learning. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1568–1576. IEEE, 2018.

[25] Bingjie Xu, Junnan Li, Yongkang Wong, Mohan S Kankanhalli, and Qi Zhao. Interact as you intend: Intention-driven human-object interaction detection. *arXiv preprint arXiv:1808.09796*, 2018.

[26] Danfei Xu, Yuke Zhu, Christopher Choy, and Li Fei-Fei. Scene graph generation by iterative message passing. In *Computer Vision and Pattern Recognition (CVPR)*, 2017.

[27] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning (ICML)*, pages 2048–2057, 2015.

[28] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. Image captioning with semantic attention. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 4651–4659, 2016.

[29] Haonan Yu, Jiang Wang, Zhiheng Huang, Yi Yang, and Wei Xu. Video paragraph captioning using hierarchical recurrent neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 4584–4593, 2016.

[30] Hanwang Zhang, Zawlin Kyaw, Shih-Fu Chang, and Tat-Seng Chua. Visual translation embedding network for visual relation detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 5532–5540, 2017.