# Action-Guided Attention Mining and Relation Reasoning Network for Human-Object Interaction Detection

**Xue Lin** , **Qi Zou** [*] and **Xixia Xu**
Beijing Key Laboratory of Traffic Data Analysis and Mining, Beijing Jiaotong University, China
{18112028, qzou, 18120432}@bjtu.edu.cn

## Abstract

Human-object interaction (HOI) detection is important to understand human-centric scenes and is challenging due to subtle difference between fine-grained actions, and multiple co-occurring interactions. Most approaches tackle the problems by considering the multi-stream information and even introducing extra knowledge, which suffer from a huge combination space and the non-interactive pair domination problem. In this paper, we propose an Action-Guided attention mining and Relation Reasoning (AGRR) network to solve the problems. Relation reasoning on human-object pairs is performed by exploiting contextual compatibility consistency among pairs to filter out the non-interactive combinations. To better discriminate the subtle difference between fine-grained actions, an action-aware attention based on class activation map is proposed to mine the most relevant features for recognizing HOIs. Extensive experiments on V-COCO and HICO-DET datasets demonstrate the effectiveness of the proposed model compared with the state-of-the-art approaches.

## 1 Introduction

Human-Object Interaction (HOI) detection task, as a sub-task of visual relationship detection, aims to localize all humans and objects, and infer the interactions between them, i.e., ⟨ human, verb, object ⟩ triplets, from an input image. HOI detection is critical for many vision tasks, such as activity analysis [Heilbron *et al.*, 2015], visual question answering (VQA) [Mallya and Lazebnik, 2016], and weakly-supervised object detection [Kim *et al.*, 2019].

However, detecting human object interaction is challenging, due to subtle differences between human-centric fine-grained actions, and multiple co-occurring interactions. Most existing works on HOI detection typically tackle the problem by combining human feature, object feature and the spatial relationship to detect HOIs. Recent approaches have attempted to improve HOI detection by integrating extra knowledge, e.g., pose cues, and word embedding. In this paper, we point out two factors which are important but neglected in different degrees.

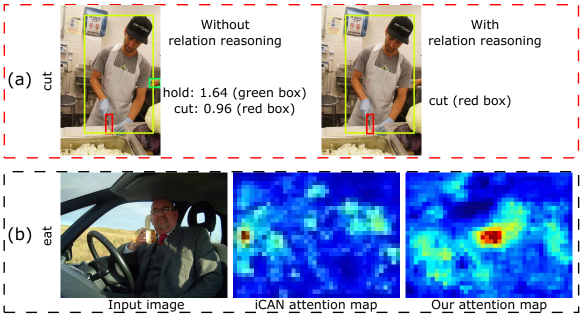

Figure 1: Illustration of our proposed model for HOI detection. (a) The model with human object pairs relation reasoning can filter out the negative combination, e.g., human and object with green box. (b) Compared with iCAN human-centric attention map, our action-aware attention map can focus on the most discriminative regions.

**First, the contextual compatibility consistency of human-object pairs is essential for accurate HOI detection.** In many works [Gao *et al.*, 2018; Wang *et al.*, 2019], detected human and objects are often paired exhaustively, which is time-consuming on some non-interactive pairs as shown in Figure 1 (a). Although [Li *et al.*, 2019c] do some efforts to eliminate some non-interactive pairs, they subject to human pose and treat human-object pairs separately with each other, without any reasoning between them. Our key insight is that human-object pairs corresponding to the accurate labels tend to have similar features, e.g., spatial structure and visual features, and hence propagating features on the relation-based graph helps learn more robust pattern of specific human-object interaction. Based on the above observation, we propose a relation reasoning model to generate enhanced visual representations of human-object pairs by leveraging the compatibility consistency of human-object pairs for filtering out the non-interactive pairs.

**Second, the human-object interaction-sensitive features should be mined and assigned more attention.** Many approaches apply the general feature extractor, e.g., ResNet-50 or ResNet-152, to obtain the small and sparse features. Specially, some methods explore the instance-centric attention [Gao *et al.*, 2018] and channel-wise and spatial atten-

[*]Corresponding author

tion [Wang *et al.*, 2019] to capture the most discriminative features. However, the obtained features are not highly action-relevant as shown in Figure 1 (b) and the action-aware features are of great importance to identify the subtle difference between actions. Studies in [Zhou *et al.*, 2016a; Selvaraju *et al.*, 2017a] suggest that class activation map (CAM) can focus on the class-aware regions and mining the attention can promote the generation ability of the model. Therefore, we are motivated to explicitly utilize the class activation map as useful information to mine the action-related attention for fine-grained interaction detection.

In this work, we propose an **A**ction-**G**uided attention mining and **R**elation **R**easoning network (**AGRR**) for human object interaction detection. The proposed framework contains human object interaction detection stream, including human/object localization and fine-grained interaction recognition, and attention mining stream. For human/object localization, we construct a directed graph whose node represents the human-object pair and edge denotes the compatibility of two neighbour nodes. Reasoning on these human-object pairs using graph attention network can enhance the pairwise features and further help filter out the irrelevant candidate pairs. Furthermore, inspired by [Li *et al.*, 2018], we introduce an action-guided attention mining loss based on the class activation map to enforce the model to learn more discriminative features for identifying the fine-grained interaction. Specially, the class activation map computed by Grad-CAM [Selvaraju *et al.*, 2017a] is human-centric for exploring the subtle difference between human actions.

To summarize, our contributions are as follows:

(1) We propose a relation reasoning model to enhance the human object pairwise features by exploiting the contextual compatibility consistency among pairs for eliminating the irrelevant candidate pairs from the exhaustive combinations.

(2) We introduce a novel action-guided attention mining loss based on the class activation map to force the model to learn more discriminative features for fine-grained interaction recognition.

(3) We perform extensive experiments on the V-COCO and HICO-DET datasets to validate the effectiveness of our proposed model, and show that it can achieve state-of-the-art results on these benchmarks.

## 2 Related Works

### 2.1 Human-Object Interaction Detection

Human-Object interaction detection is essential for understanding human activity in a complex scene. Early studies mainly focus on tackling HOIs recognition by utilizing multi-stream information, which can be divided into two categories: without and with extra knowledge.

**Without extra knowledge.** Gkioxari *et al.* [Gkioxari *et al.*, 2018] introduce an action specific density map estimation method to locate objects interacted with human. Qi *et al.* [Qi *et al.*, 2018] propose graph parsing neural network (GPNN) to model the structured scene into a graph and propagate messages between each human and object node for HOIs. Differently, we perform relation reasoning on human object pairs instead of separate human and object nodes to keep the meaningful pairs, which is more reasonable because the human object interaction is pairwise. Gao *et al.* [Gao *et al.*, 2018] and Wang *et al.* [Wang *et al.*, 2019] respectively introduce an instance-centric and contextual attention to highlight the interest region for detecting HOIs. Different from [Gao *et al.*, 2018] and [Wang *et al.*, 2019], whose attention map is human object interaction-agnostic, our proposed attention map obtained from Grad-CAM is action-aware.

**With extra knowledge.** There have been several attempts that use extra knowledge, such as word embedding [Xu *et al.*, 2019] or human pose [Fang *et al.*, 2018; Li *et al.*, 2019c; Wan *et al.*, 2019; Zhou and Chi, 2019], for detecting HOIs. Compared with [Li *et al.*, 2019c], our proposed relation reasoning model is superior in three aspects. Firstly, we only use visual features without any extra knowledge like pose to compute the similarity. Secondly, we construct a graph made up of human-object pairs rather than single human and object. Thirdly, considering the compatibility consistency, we perform relation reasoning on the relation-aware graph to obtain the enhanced features. Although the approaches using extra knowledge achieve good performance in HOI detection, they excessively rely on a well-trained model for estimating pose or extracting word embedding.

### 2.2 Relation Reasoning Methods

Relation reasoning on graph-structure data is popular and widely applied into various fields, including VQA [Li *et al.*, 2019b] and image-text matching [Li *et al.*, 2019a]. Graph Convolution Networks (GCNs) [Kipf and Welling, 2017] are proposed for semi-supervised classification. Graph Attention Networks (GATs) [Velickovic *et al.*, 2018] are then introduced to address the shortcomings of GCN that cannot deal with the structure-unknown or dynamic graph. Li *et al.* [Li *et al.*, 2019b] propose a Relation-aware Graph Attention Network (ReGAT), which learns both explicit and implicit relations between visual objects via graph attention networks, to learn question-adaptive relation representations for VQA. Li *et al.* [Li *et al.*, 2019a] propose a simple and interpretable reasoning model to generate visual representation that captures key objects and semantic concepts of a scene for image-text matching. In contrast, our method performs relation reasoning on human-object pairs rather than separate humans and objects. It is never considered before that compatibility consistency among pairs can effectively filter out irrelevant pairs.

### 2.3 Class Activation Map

Class Activation Map (CAM) [Zhou *et al.*, 2016b] is an effective tool that can highlight task-relevant regions by generating coarse class activation maps. Recently, Grad-CAM [Selvaraju *et al.*, 2017b] extends the CAM to available architectures for various tasks to provide visual explanations of model decisions. Fukui *et al.* [Fukui *et al.*, 2019] design the Attention Branch Network (ABN) for image recognition by generating the attention map for visual explanation based on CAM. Li *et al.* [Li *et al.*, 2018] propose a guided attention inference model by exploring supervision from the network itself for semantic segmentation. Inspired by the successful applications of class activation map, we propose to compute human-centric action-aware attention map based on the Grad-CAM

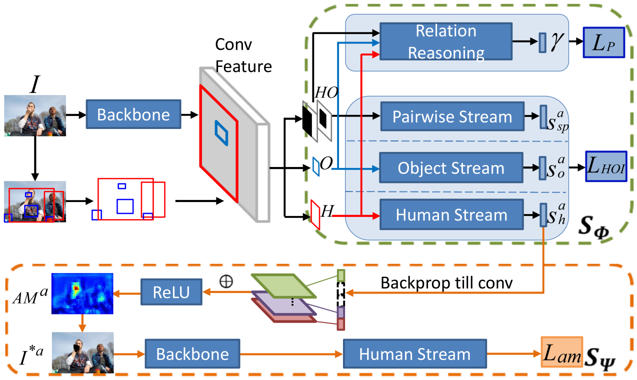

Figure 2: Overview of our framework for HOI detection comprising human object interaction detection stream $S_\Phi$ and attention mining stream $S_\Psi$. The stream $S_\Phi$ consists of a relation reasoning model (Section 3.1) for selecting the accurate pairs from the candidates and a recognition module for inferring detailed interactions. The attention mining stream $S_\Psi$ (Section 3.2), corresponding to orange lines, is only used in training phase.

rather than CAM used in [Fukui *et al.*, 2019], and guide the model to learn more discriminative features for fine-grained HOI recognition. The difference between ours and [Li *et al.*, 2018] is that we aim to find the most representative action-aware features for fine-grained interaction recognition rather than more complete areas of class.

# 3 Proposed Method

Our approach contains a human object interaction detection stream $S_\Phi$ and an attention mining stream $S_\Psi$. For $S_\Phi$, following the setting of [Gao *et al.*, 2018], we first use Faster R-CNN [Ren *et al.*, 2015] from Detectron [Girshick *et al.*, 2018] with ResNet-50-FPN [Lin *et al.*, 2017] to provide all detected human/object instances and their corresponding detection scores. Then a multi-stream architecture including human, object, and pair is built to predict interactions. Specially, we propose a relation reasoning model that can filter out irrelevant human object pairs from the infinite combinations. For $S_\Psi$, the action-aware attention map based on Grad-CAM is mined to guide the network to learn more relevant features that can discriminate the fine-grained interactions. The overall architecture is illustrated in Figure 2.

## 3.1 Human-Object Pairs Relation Reasoning

Given the detected human and object proposals, we aim to select accurate human object pairs that are actually interactive. The human and object are the indispensable components for a certain interaction, so integral consideration of human object pair is essential for HOI detection. Therefore, we perform relation reasoning on the human object pairs according to their contextual compatibility consistency to enhance the pairwise features, which can finally promote to eliminate the relatively irrelevant candidates from the infinite combinations as shown in Figure 3.

### Graph Construction

Assuming we have detected human and object bounding boxes $H = \{h_1, ..., h_m, ..., h_M\}$ and $O = \{o_1, ..., o_n, ..., o_N\}$ represented with red and blue boxes respectively as shown in Figure 3. Any human $h_m$ and any object $o_n$ will constitute a candidate human object pair $ho_i \in HO (\in \mathbb{R}^K)$. Each human $h_m$ and object $o_n$ is associated with a visual feature vector $h_m \in \mathbb{R}^{d_v}$ and $o_n \in \mathbb{R}^{d_v}$ respectively, where $d_v = 2048$ in our experiment. By treating each human object pair $ho_i$ in the image as one node $v_i \in V (\in \mathbb{R}^{K \times d_v})$, we can construct a fully-connected directed graph $G(V, E)$, where $E$ is the set of edges. Each node $v_i$ is initialized with visual features and further enhanced by neighbor nodes' information. Each edge represents the compatibility consistency whose value is high when the neighbor human-object pairs correspond to the accurate interaction and low when one pair is interactive and the other is non-interactive. The values of all edges are learned implicitly without any prior knowledge.

### Relation Reasoning Model

Inspired by recent advances in deep learning based relation reasoning [Li *et al.*, 2019b; Li *et al.*, 2019a], we perform relation reasoning on the graph $G(V, E)$ to enhance the node representations by considering the compatibility consistency among human object pairs. Specifically, we measure the pairwise affinity between human-object pairs in an embedding space to construct their relations using Eq. (1).

$$R(v_i, v_j) = \varphi(v_i)^T \phi(v_j), \tag{1}$$

where $\varphi(v_i) = W_\varphi v_i$ and $\phi(v_j) = W_\phi v_j$ mean two different embedding space. The weight parameters $W_\varphi$ and $W_\phi$ are learned by back propagation. Here, the relations between $ho_i$ and $ho_j$ are not interchangeable, meaning that the edges formed by relations are not symmetric. To make coefficients easily comparable across different nodes, we follow the routine to row-wise normalize the affinity matrix $R$ using the softmax function.

$$\alpha_{ij} = \frac{exp(R(v_i, v_j))}{\sum_k exp(R(v_i, v_k))}. \tag{2}$$

We apply Graph Attention Networks (GATs) [Velickovic *et al.*, 2018] to reason on the graph. The response of a node is computed based on its neighbors defined by the graph relations as the following attention mechanism.

$$v_i^{'} = \sigma(\sum_j \alpha_{ij} W_g v_j), \tag{3}$$

where $W_g$ is a learnable parameter with dimension of $K \times K$. $\sigma(\cdot)$ represents a nonlinear function such as ReLU. To stabilize the learning process of self-attention, we extend the above graph attention mechanism by employing multi-head attention. Specifically, $L$ independent attention mechanisms execute the transformation of Eq. (3), and then their features are concatenated as follows.

$$v_i^{'} = \|_{l=1}^{L} \sigma(\sum_j \alpha_{ij} W_g^l v_j). \tag{4}$$

In the end, $v_i^{'}$ is added to the original visual feature $v_i$ to serve as the final relation-aware enhanced features $v^*$ of human-object pairs. In order to eliminate the inaccurate HOIs, a fully-connected layer is designed as follows.

$$\gamma = sigmoid(fc(v_i^*)), \tag{5}$$

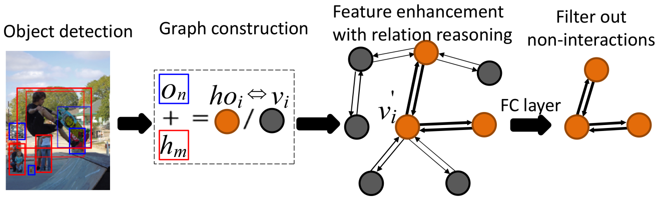
Object detection
Graph construction
Feature enhancement with relation reasoning
Filter out non-interactions
$O_n$
$h_m$
$ho_i \Leftrightarrow v_i$
+ =
/
$v_i'$
FC layer

Figure 3: The diagram of relation reasoning on human object pairs. For detected humans (red square boxes) and objects (blue square boxes) in an input image, any human and any object are combined exhaustively to form the candidate HOI pairs, including the correct HOIs (orange circle) and incorrect HOIs (black circle). The human-object pairs that correspond to accurate HOIs are highly consistent and connected with thick arrows. Then relation reasoning is performed on the graph to obtain the enhanced human-object pairwise features for disregarding the irrelevant pairs.

where $\gamma \in \mathbb{R}^2$ represents whether the $ho_i$ pair corresponding to $v_i^*$ is correct or not. Relation reasoning on the human-object pairs can filter out the non-interactive pairs from all infinite candidates by considering the contextual information, which eventually improves the HOI detection performance.

## 3.2 Action-aware Attention Mining

To make our model learn the discriminative features for fine-grained interaction detection, we explore the class activation map from the network itself to mine the action-aware features and further introduce a novel attention mining loss to guide the model to learn effectively.

### Class Activation Map for HOIs

In Figure 2, the attention mining stream $S_\Psi$, sharing parameters with human object interaction detection stream $S_\Phi$, aims to find out the discriminative regions that are beneficial to interactions detection. Based on the fundamental framework of Grad-CAM, we generate the attention map of human-centric interaction. In stream $S_\Phi$, for a given image $I$, let $F_i$ be the activation of features maps $i$ in the $4$-$th$ block layer. Class specific attention maps can be obtained by computing the gradient of the score $s_h^a$ for interaction class $a$ of human stream, with respect to activation maps $F_i(x, y)$. A global average pooling operation is then performed on these gradients to obtain the importance weights $w_i^a$ as follows.

$$w_i^a = \frac{1}{\theta * \beta} \sum_{x,y} \frac{\partial s_h^a}{\partial F_i(x,y)}, \tag{6}$$

where $\theta, \beta$ mean the width and height of feature map $F_i$.

Following the works [Selvaraju *et al.*, 2017a; Li *et al.*, 2018], the class attention map $AM^a$ is a weighted combination of forward activation maps $F$ followed by a ReLU as follows.

$$AM^a = ReLU(\sum_i w_i^a F_i), \tag{7}$$

where the ReLU function is applied to the linear combination of feature maps because we are only interested in the features that have a positive influence on the interaction.

### Attention Mining

Based on the above class attention map, we design a mask to be applied on the original input image using Eq. (8).

$$I^{*a} = I - (T(AM^a) \odot I), \tag{8}$$

where $\odot$ denotes element-wise multiplication. $T(AM^a)$ is a masking function based on a threshold as follows.

$$T(AM^a) = \begin{cases} 0 & AM^a(x,y) < t \\ 1 & AM^a(x,y) > t, \end{cases} \tag{9}$$

where $t$ is a threshold equal to the median of maximum and minimum values of $AM^a$ for binarizing the attention map.

The masked image $I^{*a}$ is then used as input of the attention mining stream $S_\Psi$ to obtain the interaction class prediction score. Since our goal is to guide the network to focus on the action-aware representative features, we enforce $I^{*a}$ to contain as little features belonging to the target action as possible. With respect to the loss function, the model tries to minimize the prediction score of $I^{*a}$ for interaction class $a$. Thus, we introduce an attention mining loss as defined in Eq. (10).

$$L_{am} = \frac{1}{Z} \sum_c s_h^a(I^{*a}), \tag{10}$$

where $s_h^a(I^{*a})$ represents the prediction score of $I^{*a}$ for interaction class $a$. $Z$ is the number of ground-truth interaction labels for image $I$.

## 3.3 Inference and Training

**Inference.** For each human-object pair $ho_i$, we first decide whether the pair is interactive or not according to $\gamma$ obtained from the relation reasoning model. Non-interactive pairs are excluded, and the score $s_{h,o}^a$ for each accurate interaction is predicted. Following [Gao *et al.*, 2018; Wang *et al.*, 2019], the score $s_{h,o}^a$ depends on (1) the confidence for the individual object detections $s_h$ and $s_o$, (2) the interaction prediction based on the appearance of the person $s_h^a$ and object $s_o^a$, and (3) the score prediction based on the spatial relationship between the person and object $s_{sp}^a$. Specifically, our HOI score $s_{h,o}^a$ can be formulated as.

$$s_{h,o}^a = s_h \cdot s_o \cdot (s_h^a + s_o^a) \cdot s_{sp}^a. \tag{11}$$

**Multi-task Training.** Since a person can concurrently perform different actions to one or multiple target objects, HOI detection is thus a multi-label classification problem. We apply binary sigmoid classifier for each action category, and minimize the binary cross entropy losses between action scores $s_h^a$, $s_o^a$, $s_{sp}^a$ and the ground-truth action labels for each action category, denoted as $L_{HOI}$ that is generally summed over the above losses with weight of one, except for the loss term for $s_h^a$ with a weight of two. In addition, we minimize the binary cross entropy losses between $\gamma$ (obtained from Eq. (5)) and the ground-truth interactive labels for each pair, denoted as $L_P$. Our overall loss is summed over all losses as follows.

$$L = L_{HOI} + L_P + \eta * L_{am}, \tag{12}$$

where $\eta$ is a hyper-parameter that controls the importance of attention mining loss. We use $\eta = 2$ in all of our experiments in order to coincide to the weight of loss term for $s_h^a$ in $L_{HOI}$.

| Methods | Extra knowledge | Feature Backbone | Default | | | Known Object | | |
|---|---|---|---|---|---|---|---|---|
| | | | Full | Rare | Non-Rare | Full | Rare | Non-Rare |
| Xu *et al.* [Xu *et al.*, 2019] | Word embedding | ResNet-50 | 14.70 | 13.26 | 15.13 | - | - | - |
| RPNN [Zhou and Chi, 2019] | Pose | ResNet-50 | 17.35 | 12.78 | 18.71 | - | - | - |
| Li *et al.* ($RP_DC_D$) [Li *et al.*, 2019c] | Pose | ResNet-50 | 17.03 | 13.42 | 18.11 | 19.17 | 15.51 | 20.26 |
| PMFNet [Wan *et al.*, 2019] | Pose | ResNet-50-FPN | 17.46 | 15.65 | 18.00 | 20.34 | 17.47 | 21.20 |
| InteractNet [Gkioxari *et al.*, 2018] | None | ResNet-50-FPN | 9.94 | 7.16 | 10.77 | - | - | - |
| GPNN [Qi *et al.*, 2018] | None | ResNet-152 | 13.11 | 9.34 | 14.23 | - | - | - |
| iCAN [Gao *et al.*, 2018] | None | ResNet-50 | 14.84 | 10.45 | 16.15 | 16.26 | 11.33 | 17.73 |
| Wang *et al.* [Wang *et al.*, 2019] | None | ResNet-50 | 16.24 | 11.16 | 17.75 | 17.73 | 12.78 | 19.21 |
| Ours | None | ResNet-50 | **16.63** | **11.30** | **18.22** | **19.22** | **14.56** | **20.61** |

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

| Methods | Extra knowledge | Feature Backbone | mAP$_{role}$ |
|---|---|---|---|
| Xu *et al.* [Xu *et al.*, 2019] | Word embedding | ResNet-50 | 45.9 |
| RPNN [Zhou and Chi, 2019] | Pose | ResNet-50 | 47.5 |
| Li *et al.* ($RP_DC_D$) [Li *et al.*, 2019c] | Pose | ResNet-50 | 47.8 |
| PMFNet [Wan *et al.*, 2019] | Pose | ResNet-50-FPN | 52 |
| Gupta *et al.* [Gupta and Malik, 2015] | None | ResNet-50-FPN | 31.8 |
| InteractNet [Gkioxari *et al.*, 2018] | None | ResNet-50-FPN | 40.0 |
| GPNN [Qi *et al.*, 2018] | None | ResNet-152 | 44.0 |
| iCAN [Gao *et al.*, 2018] | None | ResNet-50 | 45.3 |
| Wang *et al.* [Wang *et al.*, 2019] | None | ResNet-50 | 47.3 |
| Ours | None | ResNet-50 | **48.1** |

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

# 4 Experimental Results

## 4.1 Datasets and Metrics

**Datasets.** To verify the effectiveness of our method, we conduct experiments on two HOI benckmark datasets, i.e., V-COCO [Gupta and Malik, 2015] and HICO-DET [Chao *et al.*, 2018] datasets. V-COCO is a subset of MS-COCO [Lin *et al.*, 2014], 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. Note that three action classes (i.e., cut, hit, eat) are annotated with two types of targets (i.e., instrument and direct object). HICO-DET [Chao *et al.*, 2018] 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 [Lin *et al.*, 2014]) and 117 unique action verbs.

**Evaluation Metrics.** Following the standard evaluation setting in [Gao *et al.*, 2018], we use role mean average precision (mAP) to measure the HOI detection performance. The goal is to detect the agents and the objects in the various roles for the action. The HOI detection is considered as a true positive if it has the correct action label, and the intersection-over union (IoU) between the human and object bounding-box predictions and the respective ground-truth boxes is greater than the threshold 0.5.

## 4.2 Implementation Details

We deploy Detectron [Girshick *et al.*, 2018] with a ResNet-50-FPN [Lin *et al.*, 2017] backbone to obtain human and object bounding-box predictions. To select a predicted bounding-box as a training sample, we set the confidence threshold to be 0.8 for humans and 0.4 for objects. For fair comparison, we adopt the object detection results and pre-trained weights from [Gao *et al.*, 2018]. The low-grade instance suppressive training strategy [Li *et al.*, 2019c] is also applied. We use SGD optimizer for training with initial learning rate 5e-6, weight decay 5e-4 and a momentum 0.9 for all datasets. In training, the ratio of positive and negative samples is 1:3. All experiments are conducted on a single Nvidia Titan XP GPU.

## 4.3 Quantitative Results

We compare our proposed model with several existing approaches trained with and without extra knowledge. For V-COCO dataset, we evaluate mAP$_{role}$ of 24 actions with roles as in work [Gupta and Malik, 2015]. As shown in Table 1, our method achieves 48.1 mAP$_{role}$, outperforming all existing approaches without extra knowledge. Compared with the methods with extra knowledge except PMFNet [Wan *et al.*, 2019], ours is more superior, which demonstrates our model can learn to capture the subtle difference of fine-grained interactions even without any extra knowledge.

For HICO-DET dataset, we report results on three different HOI category sets: full, rare, and non-rare with two different settings of Default and Known Objects [Xu *et al.*, 2019]. As shown in Table 2, ours outperforms the state-of-the-arts without extra knowledge in all settings. Specially, our proposed model achieves great performance with 16.63 mAP and 19.22 mAP on Default and Know Object categories respectively, with relative gains of 0.39 and 1.49 over the best existing method [Wang *et al.*, 2019]. In addition, compared with [Xu *et al.*, 2019] trained using word embedding, ours achieves 1.93 mAP improvements in full category sets under Default settings. Extra knowledge usually can bring gains because they provide the guided signal to capture fine-grained interaction features. However, they extremely rely on the pre-trained pose estimation model that also needs a large number of labeled samples to train.

## 4.4 Ablation Study

**Effectiveness of Human-Object Pairs Relation Reasoning.** As shown in Table 3, the relation reasoning model achieves a significant mAP$_{role}$ improvement from 45.18 to 47.74 on V-COCO dataset. For HICO-DET dataset, as shown in Table 4, it improves the mAP by 0.51 and 0.57 over the baseline on full category with two different settings of Default and Known Objects. It demonstrates that our proposed relation

| Methods | w/ human-object pairs relation reasoning | w/ attention mining | mAP$_{role}$ |
|---|---|---|---|
| Baseline | - | - | 45.18 |
| | √ | - | 47.74 |
| | - | √ | 47.32 |
| | √ | √ | 48.10 |

Table 3: Ablation study on the V-COCO dataset about human-object pairs relation reasoning and attention mining.

| Models | Default | | | Known Object | | |
|---|---|---|---|---|---|---|
| | Full | Rare | Non-Rare | Full | Rare | Non-Rare |
| Baseline | 15.67 | 10.37 | 17.25 | 18.14 | 13.28 | 19.59 |
| w/ human-object pairs relation reasoning | 16.18 | 10.14 | 17.98 | 18.71 | 13.19 | 20.36 |
| w/ attention mining | 16.45 | 10.92 | 18.10 | 19.05 | 14.21 | 20.50 |
| w/ all | 16.63 | 11.30 | 18.22 | 19.22 | 14.56 | 20.61 |

Table 4: Ablation study on the HICO-DET dataset about human-object pairs relation reasoning and attention mining.

| Methods | Directed edges | Multi-head | mAP$_{role}$ |
|---|---|---|---|
| w/o relation reasoning | - | - | 45.18 |
| | √ | - | 47.51 |
| | - | √ | 47.71 |

Table 5: Ablation study on the V-COCO dataset about the directed edges and multi-head attention of relation reasoning.

reasoning model can enhance the human-object pairs features and further benefit to the elimination of non-interactive pairs.

**Effectiveness of Action-aware Attention Mining.** The action-aware attention mining model aims to mine the most representative features by exploring the action-aware attention for recognizing fine-grained interactions. It can be observed from Table 3 that it exceeds the baseline 2.14 at metric mAP$_{role}$ on V-COCO dataset. For HICO-DET dataset, it achieves 0.78 and 0.91 mAP$_{role}$ improvements on Default and Known Objects settings compared with the baseline as shown in Table 4. This strongly indicates that the action-aware attention mining can force the model to learn more discriminative features for HOI detection.

**Effectiveness of directed edges and multi-head attention.** We can observe from Table 5 that the model with directed edges achieve 2.33 mAP$_{role}$ improvement compared with the baseline. In addition, the model with multi-head attention improves the mAP$_{role}$ performance of baseline by 2.53. It reveals that both directed edge and multi-head attention can significantly improve the performance of HOI detection.

### 4.5 Qualitative Visualization Results

Figure 4 shows interaction detection examples that the detected human has different interactions with various objects simultaneously. Figure 5 (a) displays the HOI detections on V-COCO test set [Gupta and Malik, 2015], which demonstrates that our model can detect various objects that the human instances are interacting with in different situations. Figure 5 (b) represents the sample HOI detections on HICO-DET test set [Chao *et al.*, 2018]. It demonstrates that our approach can detect multiple interactions with the same object.

Figure 6 (a) reveals that our proposed relation reasoning model can effectively disregard the irrelevant pairs from the numerous candidates, especially suitable for the case that multiple persons have different interactions. Figure 6 (b) indicates that our proposed action-guided attention map can extract the more discriminative features than iCAN [Gao *et al.*, 2018] for detecting fine-grained interactions.

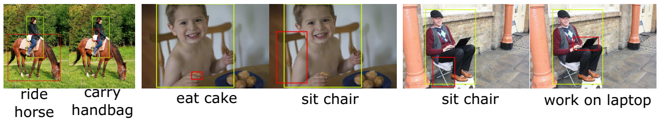

Figure 4: Multiple interaction detections on V-COCO dataset. Our model detects the human instance having different interactions with different objects.

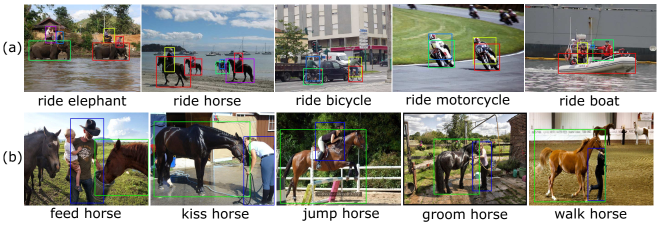

Figure 5: (a) Sample HOI detections on the V-COCO dataset. Our model detects various objects that the human instances are interacting with in different situations. (b) Sample HOI detections on the HICO-DET dataset. Our model detects different types of interactions with objects from the same category.

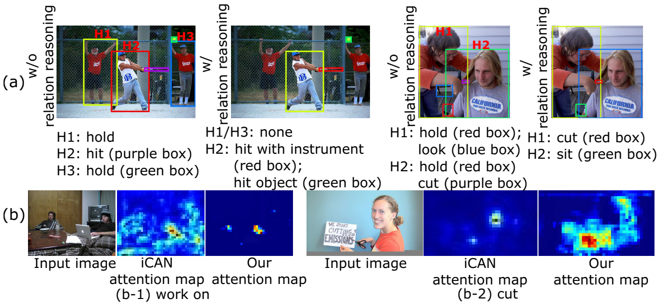

Figure 6: (a) Influence of relation reasoning on filtering out the inaccurate combinations. (b) Comparison of attention maps obtained using our approach and iCAN model.

## 5 Conclusion

In this paper, we propose a novel method for HOI detection. Our approach performs relation reasoning on human-object pairs by exploring contextual compatibility consistency among pairs to disregard the irrelevant combinations. Furthermore, we introduce an action-guided attention mining loss to achieve the fine-grained interaction detection. The experimental results demonstrate our proposed method can achieve a comparable performance with the state-of-the-art methods.

## Acknowledgments

This research is supported by National Nature Science Foundation of China (No.61473031), the Fundamental Research Funds for the Central Universities (2019YJS030, 2019JBZ104).

## References

[Chao *et al.*, 2018] Yuwei Chao, Yunfan Liu, Xieyang Liu, Huayi Zeng, and Jia Deng. Learning to detect human-object interactions. In *WACV*, pages 381–389, 2018.

[Fang *et al.*, 2018] Haoshu Fang, Jinkun Cao, Yuwing Tai, and Cewu Lu. Pairwise body-part attention for recognizing human-object interactions. In *ECCV*, pages 52–68, 2018.

[Fukui *et al.*, 2019] Hiroshi Fukui, Tsubasa Hirakawa, Takayoshi Yamashita, and Hironobu Fujiyoshi. Attention branch network: Learning of attention mechanism for visual explanation. In *CVPR*, pages 10705–10714, 2019.

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

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

[Gkioxari *et al.*, 2018] Georgia Gkioxari, Ross Girshick, Piotr Dollar, and Kaiming He. Detecting and recognizing human-object interactions. In *CVPR*, pages 8359–8367, 2018.

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

[Heilbron *et al.*, 2015] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *CVPR*, pages 961–970, 2015.

[Kim *et al.*, 2019] Daesik Kim, Gyuejeong Lee, Jisoo Jeong, and Nojun Kwak. Tell me what they're holding: Weakly-supervised object detection with transferable knowledge from human-object interaction. *arXiv:1911.08141*, 2019.

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

[Li *et al.*, 2018] Kunpeng Li, Ziyan Wu, Kuanchuan Peng, Jan Ernst, and Yun Fu. Tell me where to look: Guided attention inference network. In *CVPR*, pages 9215–9223, 2018.

[Li *et al.*, 2019a] Kunpeng Li, Yulun Zhang, Kai Li, Yuanyuan Li, and Yun Fu. Visual semantic reasoning for image-text matching. In *ICCV*, 2019.

[Li *et al.*, 2019b] Linjie Li, Zhe Gan, Yu Cheng, and Jingjing Liu. Relation-aware graph attention network for visual question answering. In *ICCV*, 2019.

[Li *et al.*, 2019c] 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. In *CVPR*, 2019.

[Lin *et al.*, 2014] Tsungyi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollar, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755, 2014.

[Lin *et al.*, 2017] Tsungyi Lin, Piotr Dollar, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *CVPR*, pages 936–944, 2017.

[Mallya and Lazebnik, 2016] Arun Mallya and Svetlana Lazebnik. Learning models for actions and person-object interactions with transfer to question answering. In *ECCV*, pages 414–428, 2016.

[Qi *et al.*, 2018] Siyuan Qi, Wenguan Wang, Baoxiong Jia, Jianbing Shen, and Songchun Zhu. Learning human-object interactions by graph parsing neural networks. In *ECCV*, pages 407–423, 2018.

[Ren *et al.*, 2015] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: towards real-time object detection with region proposal networks. 2015:91–99, 2015.

[Selvaraju *et al.*, 2017a] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*, pages 618–626, 2017.

[Selvaraju *et al.*, 2017b] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*, pages 618–626, 2017.

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

[Wan *et al.*, 2019] Bo Wan, Desen Zhou, Yongfei Liu, Rongjie Li, and Xuming He. Pose-aware multi-level feature network for human object interaction detection. In *ICCV*, 2019.

[Wang *et al.*, 2019] Tiancai Wang, Rao Muhammad Anwer, Muhammad Haris Khan, Fahad Shahbaz Khan, Yanwei Pang, Ling Shao, and Jorma Laaksonen. Deep contextual attention for human-object interaction detection. In *ICCV*, 2019.

[Xu *et al.*, 2019] Bingjie Xu, Yongkang Wong, Junnan Li, Qi Zhao, and Mohan S Kankanhalli. Learning to detect human-object interactions with knowledge. In *CVPR*, pages 2019–2028, 2019.

[Zhou and Chi, 2019] Penghao Zhou and Mingmin Chi. Relation parsing neural network for human-object interaction detection. In *ICCV*, pages 843–851, 2019.

[Zhou *et al.*, 2016a] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *CVPR*, pages 2921–2929, 2016.

[Zhou *et al.*, 2016b] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *CVPR*, pages 2921–2929, 2016.