Transferable Interactiveness Knowledge for Human-Object Interaction Detection

Yong-Lu Li, Siyuan Zhou, Xijie Huang, Liang Xu, Ze Ma, Hao-Shu Fang, Yan-Feng Wang, Cewu Lu*
Shanghai Jiao Tong University
{yonglu_li, ssluvble, otaku_huang, liangxu, maze1234556}@sjtu.edu.cn
fhaoshu@gmail.com, wangyanfeng@sjtu.edu.cn, lucewu@sjtu.edu.cn

Abstract

Human-Object Interaction (HOI) Detection is an important problem to understand how humans interact with objects. In this paper, we explore ***Interactiveness Knowledge*** *which indicates whether human and object interact with each other or not. We found that interactiveness knowledge can be learned across HOI datasets, regardless of HOI category settings. Our core idea is to exploit an Interactiveness Network to learn the general interactiveness knowledge from multiple HOI datasets and perform Non-Interaction Suppression before HOI classification in inference. On account of the generalization of interactiveness, interactiveness network is a transferable knowledge learner and can be cooperated with any HOI detection models to achieve desirable results. We extensively evaluate the proposed method on HICO-DET and V-COCO datasets. Our framework outperforms state-of-the-art HOI detection results by a great margin, verifying its efficacy and flexibility. Code is available at* https://github.com/DirtyHarryLYL/Transferable-Interactiveness-Network.

1. Introduction

Human-Object Interaction (HOI) detection retrieves human and object locations and infers the interaction classes from still image. As a sub-task of visual relationship [24, 17], HOI is strongly related to the human body and object understanding [33, 36, 39, 11, 26, 21, 38]. It is crucial for behavior understanding and can facilitate activity understanding [9, 28], imitation learning [3], etc. Recently, impressive progress has been made by utilizing Deep Neural Networks (DNNs) in this area [34, 19, 32, 31].

*Cewu Lu is the corresponding author, he is also a member of Department of Computer Science and Engineering, Shanghai Jiao Tong University, MoE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University, and SJTU-SenseTime AI lab.

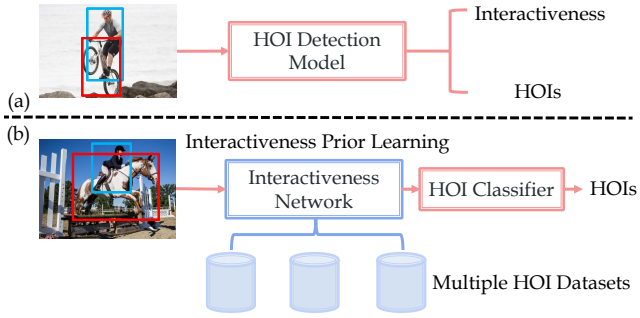

Figure 1. Interactiveness Knowledge Learning. (a) HOI datasets contain implicit interactiveness knowledge. We can learn it better by performing explicit interactiveness discrimination, and utilize it to improve the HOI detection performance. (b) Interactiveness knowledge is beyond the HOI categories and can be learned across datasets, which can bring greater performance improvement.

Generally, human and objects need to be detected first. Given an image and its detections, human and objects are often paired exhaustively [19, 31, 32]. HOI detection task aims to classify these pairs as different HOI categories. Previous one-stage methods [34, 19, 31, 13, 32] directly classify a pair as specific HOIs. These methods actually predict *interactiveness* implicitly at the same time, where interactiveness indicates whether a human-object pair is interactive. For example, when a pair is classified as HOI "eat apple", we can implicitly predict that it is interactive.

Though interactiveness is an essential element for HOI detection, we neglected to study how to utilize it and improve its learning. In comparison to HOI categories, interactiveness conveys more basic information. Such attribute makes it easier for interactiveness to transfer across datasets.Based on this inspiration, we propose a **Interactiveness Knowledge** learning method as seen in Figure 1. With our framework, interactiveness can be learned across datasets and applied to any specific dataset. By utilizing interactiveness, we take two stages to identify HOIs: we first discriminate a human-object pair as interactive or not

and then classify it as specific HOIs. Compared to previous one-stage method [34, 19, 31, 13, 32], we take advantage of powerful interactiveness knowledge that incorporates more information from other datasets. Thus our method can decrease the false positives significantly. Additionally, after the interactiveness filtering in the first stage, we do not need to handle a large number of non-interactive pairs which are overwhelmingly more than interactive ones.

In this paper, we proposed a novel two-stage method to classify pairs hierarchically as shown in Figure 2. We introduce an interactiveness network which can be combined with any HOI detection model. We set a hierarchical logical strategy: by utilizing binary interactiveness labels, interactiveness network will bring in a strong supervised constraint which refines the framework in training and learns the interactiveness from multiple datasets. In testing, interactiveness network performs Non-Interaction Suppression (NIS) first. Then the HOI detection model will classify the remaining pairs as specific HOIs, where non-interactive pairs have been decreased significantly. Moreover, if the model classifies a pair as specific HOIs, it should figure out that the pair is interactive simultaneously. Such two-stage prediction will alleviate the learning difficulty and bring in hierarchical predictions. For special attention, interactiveness offers extra information to help HOI classification and is independent of HOI category settings. That means it can be transferred across datasets and utilized to enhance HOI models designed for different HOI settings.

We perform extensive experiments on HICO-DET [34], V-COCO [13] datasets. Our method cooperated with transferred interactiveness outperforms the state-of-the-art methods by **2.38**, **3.06**, and **2.17** mAP on three Default category sets on HICO-DET, **4.0** and **3.4** mAP on V-COCO.

2. Related Works

Visual Relationship Detection. Visual relationship detection [6, 17, 24, 16] aims to detect the objects and classify their relationships simultaneously. In [17], Lu *et al*. proposed a relationship dataset VRD and an approach combined with language priors. Predicates within relationship triplet $\langle subject, predicate, object \rangle$ include actions, verbs, spatial and preposition vocabularies. Such vocabulary setting and severe long-tail issue within the dataset make this task quite difficult. Large-scale dataset Visual Genome [24] is then proposed to promote studies in this problem. Recent works [23, 25, 40, 30] put attention on more effective and efficient visual feature extraction and try to exploit semantic information to refine the relationship detection.

Human-Object Interaction Detection. Human-Object Interaction [1, 4, 2] is essential to understand human-centric interaction with objects. Recently several large-scale datasets, such as V-COCO [13], HICO-DET [34], HCVRD [18], were proposed for the exploration of HOI

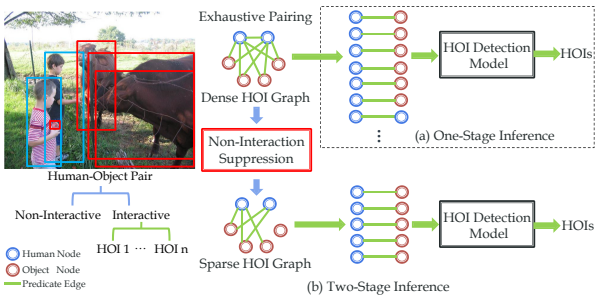

Figure 2. HOIs within an image can be represented as a HOI graph. Human and object can be seen as nodes, whilst the interactions are represented as edges. Exhaustive pairing of all nodes would import overmuch non-interactive edges and do damage to detection performance. Our Non-Interaction Suppression can effectively reduce non-interactive pairs. Thus the dense graph would be converted to a sparse graph and then be classified.

detection. Different from HOI recognition [35, 5, 12, 8, 15] which is an image level classification problem, HOI detection needs to detect interactive human-object pairs and classify their interactions at instance level. With the assistance of DNNs and large-scale datasets, recently methods have made significant progress. Chao *et al*. [34] proposed a multi-stream model combining visual features, spatial locations to help tackle this problem. To address the long tail issue, Shen *et al*. [37] studied zero-shot learning problem and predicted the verb and object separately. In [19], an action specific density map estimation method is introduced to locate objects interacted with human. In [32], Qi *et al*. proposed GPNN incorporating DNN and graphical model, which uses message parsing to iteratively update states and classifies all possible pairs/edges. Gao *et al*. [31] exploited an instance centric attention module to enhance the information from the interest region and facilitate the HOI classification. Generally, these methods inference in one-stage and may suffer from severe non-interactive pair domination problem. To address this issue, we utilize interactiveness to explicitly discriminate non-interactive pairs and suppress them before HOI classification.

3. Preliminary

HOI representation can be described as a graph model [32, 23] as seen in Figure 2. Instances and relations are expressed as nodes and edges respectively. With exhaustive pairing [19, 31], HOI graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ is dense connected, where $\mathcal{V}$ includes human node $\mathcal{V}_h$ and object node $\mathcal{V}_o$. Let $v_h \in \mathcal{V}_h$ and $v_o \in \mathcal{V}_o$ denote the human and object nodes. Thus edges $e \in \mathcal{E}$ are expressed as $e = (v_h, v_o) \in \mathcal{V}_h \times \mathcal{V}_o$. With n nodes, exhaustive paring will generate a mass of edges. We aim to assign HOI (including no HOI) labels on those edges. Considering that a vast majority of non-interactive edges existing in $\mathcal{E}$ should be discarded, our goal is to seek a sparse $\mathcal{G}^*$ with corrected

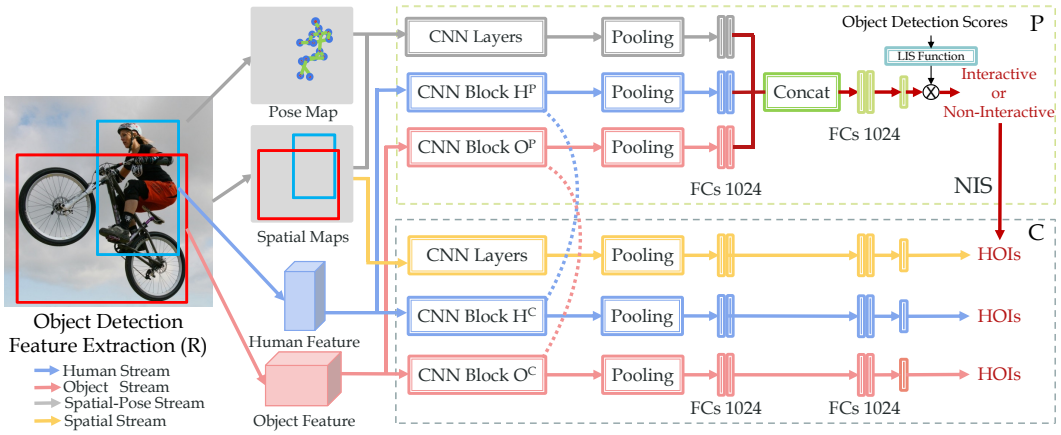

Figure 3. Overview of our framework. Interactiveness network $\mathbf{P}$ can cooperate with any HOI models (referred as $\mathbf{C}$). $\mathbf{P}$ employs human, object and spatial-pose streams to extract features from human and object appearance, spatial locations and human pose information. The outputs of three streams are concatenated and inputted to the interactiveness discriminator. When cooperated with multi-stream $\mathbf{C}$ such as [34, 31] (human, object, and spatial streams), $H^{\mathbf{P}}$ and $O^{\mathbf{P}}$ in $\mathbf{P}$ can share weights (dotted lines) with $H^{\mathbf{C}}$ and $O^{\mathbf{C}}$ in $\mathbf{C}$ during joint training. In this work, these four blocks are all residual blocks [14]. LIS and NIS will be detailed in Section 4.3 and Section 4.5.

HOI labeling on its edges.

4. Our Method

4.1. Overview

As aforementioned, we introduce **Interactiveness Knowledge** to advance HOI detection performance. That is, explicitly discriminate the non-interactive pairs and suppress them before HOI classification. From the semantic point of view, interactiveness provides more general information than conventional HOI categories. Since any human-object pair can be assigned binary interactiveness labels according to the HOI annotations, *i.e.* "interactive" or "non-interactive", interactiveness knowledge can be learned from multiple datasets with different HOI category settings and transferred to any specific datasets.

To exploit this cue, we proposed interactiveness network (interactiveness predictor, referred as $\mathbf{P}$) which utilizes interactiveness to reduce false positives caused by overmuch non-interactive pair candidates. Some conventional modules are also included, namely, Representation Network $\mathbf{R}$ (feature extractor) and Classification Network $\mathbf{C}$ (HOI classifier). $\mathbf{R}$ is responsible for feature extraction from detected instances. $\mathbf{C}$ utilizes node and edge features to perform HOI classification. Figure 3 is an overview of our framework which follows the hierarchical classification paradigm. Specifically, we first train $\mathbf{P}$ and $\mathbf{C}$ jointly to learn the interactiveness and HOIs knowledge. Under usual circumstances, the ratio of non-interactive edges is dominant within inputs. Hence $\mathbf{P}$ will bring a strong supervised signal to refine the framework. In testing, $\mathbf{P}$ is utilized in two stages. First, $\mathbf{P}$ evaluates the interactiveness of edges by exploiting the learned interactiveness knowledge, so we can convert the dense HOI graph to a sparse one. Second, combined with interactiveness score from $\mathbf{P}$, $\mathbf{C}$ will process the sparse graph and classify the remaining edges.

In addition, on account of the generalization ability of interactiveness knowledge, it can be transferred with $\mathbf{P}$ across datasets (Section 4.4). Details of the framework architecture are illustrated in Section 4.2 and 4.3. The process of training and testing will be detailed in Section 4.4.

4.2. Representation and Classification Networks

Human and Object Detection. In HOI detection, human and object need to be detected first. In this work, we follow the setting of [31] and employ the Detectron [29] with ResNet-50-FPN [20] to prepare bounding boxes and detection scores. Before post-processing, detection results will be filtered by the detection score thresholds first.

Representation Network. In previous methods [34, 19, 31], $\mathbf{R}$ is often modified from object detector such as Fast R-CNN [10] or Faster R-CNN [11]. We also exploited a Faster R-CNN [11] with ResNet-50 [14] based $\mathbf{R}$ here. During training and testing, $\mathbf{R}$ is frozen and acts as a feature extractor. Given the detected bounding boxes, we produce human and object features by cropping ROI pooling feature maps according to box coordinates.

HOI Classification Network. As for $\mathbf{C}$, multi-stream architecture and late fusion strategy are frequently used and approved effective [34, 31]. Follow [34, 31], for our classification network $\mathbf{C}$, we utilize a human stream and an object stream to extract human, object and context features. Within each stream, a residual block [14] (denoted as H^C, O^C, seen in Figure 3) with pooling layer and fully connected layers (FCs) are adopted. Moreover, an extra spa-

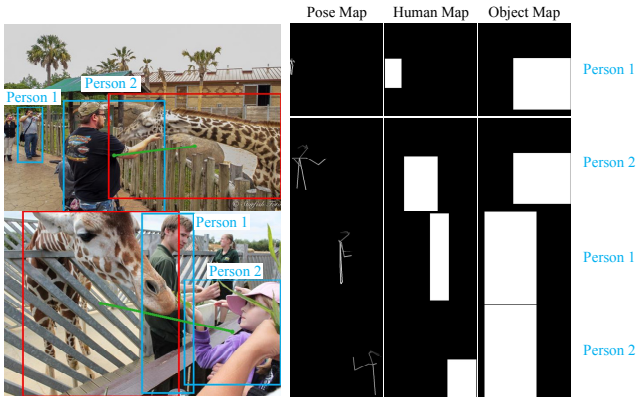

Figure 4. Inputs of the spatial-pose stream. Three kinds of maps are included: pose map, human map and object map. Person 2 in two images both have interaction "feed" with giraffes. But two pairs of Person 1 and giraffe are all non-interactive. Their poses and locations are helpful for the interactiveness discrimination.

tial stream [34] is adopted to encode the spatial locations of instances. Its input is a two-channel tensor consisting of a human map and an object map, shown in Figure 4. Human and object maps are all 64x64 and obtained from the human-object union box. In the human channel, the value is 1 in the human bounding box and 0 in other areas. The object channel is similar which has value 1 in the object bounding box and 0 elsewhere. Following the late fusion strategy, each stream will first perform HOI classification, then three prediction scores will be fused by element-wise sum in the same proportion to produce the final result of $\mathbf{C}$.

4.3. Interactiveness Network

Interactiveness needs to be learned by extracting and combining essential information. The visual appearance of human and object are obviously required. Besides, interactive and non-interactive pairs also have other distinguishing features, e.g. spatial location and human pose information. For example, in the upper image of Figure 4, Person 1 and the giraffe far from him are not interactive. Their spatial maps [34] can provide pieces of evidence to help with classification. Furthermore, pose information is also helpful. In the lower image, although two people are both close to the giraffe, only Person 2 and the giraffe are interactive. The arm of Person 2 is uplift and touching the giraffe. Whilst Person 1 is back on to the giraffe, and his pose is quite different from the typical pose of "feed".

Based on these reasons, the combination of visual appearance, spatial location and human pose information is key to interactiveness discrimination. Hence $\mathbf{P}$ needs to encode these key elements together to learn the interactiveness knowledge. A natural choice is the multi-stream architecture as presented: human, object and spatial-pose streams.

Human and Object stream. For human and object appearance, we extract ROI pooling features from representation network $\mathbf{R}$, then input them into residual blocks $H^{\mathbf{P}}$ and $O^{\mathbf{P}}$, respectively. The architecture of $H^{\mathbf{P}}$ and $O^{\mathbf{P}}$ are same as $H^{\mathbf{C}}$ and $O^{\mathbf{C}}$ (Figure 3). Through subsequent global average pooling and FCs, the output features of two streams are denoted as f_h and f_o, respectively.

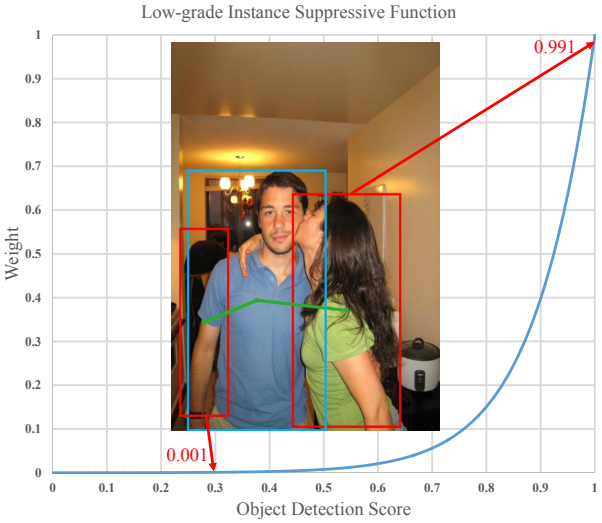

Figure 5. The illustration of $\mathcal{P}(\cdot)$ within Low-grade Suppressive Function. Its input is object detection score. High-grade detected objects will be emphasized and distinguished with low-grade ones. In addition, $\mathcal{P}(0) = 5.15E-05$ and $\mathcal{P}(1) = 9.99E-01$.

Spatial-Pose Stream. Different from [34], our spatial-pose stream input includes a special 64x64 pose map. Given the union box of each human and his/her paired object, we employ pose estimation [22, 27] to estimate his/her 17 keypoints (in COCO format [7]). Then, we link the keypoints with lines of different gray value ranging from 0.15 to 0.95 to represent different body parts, which implicitly encodes the pose features. Whilst the other area is set as 0. Finally, we reshape the union box to 64x64 to construct the pose map. We concatenate the pose map with human and object maps which are the same as those in the spatial stream of $\mathbf{C}$. This forms the input for our spatial-pose stream. Next, we exploit two convolutional layers with max pooling and two 1024 sized FCs to extract the feature f_{sp} of three maps. Last, the output will be concatenated with the outputs of human and object streams for interactiveness discrimination.

Given a HOI graph $\mathcal{G}$ with all possible edges, $\mathbf{P}$ will evaluate the interactiveness of pair (v_h, v_o) based on learned knowledge, and gives confidence:

$$s^P_{(h,o)} = f_{\mathbf{P}}(f_h, f_o, f_{sp}) * L(s_h, s_o), \tag{1}$$

where $L(s_h, s_o)$ is a novel weight function named Low-grade Instance Suppressive Function (LIS). It takes the human and object detection scores s_h, s_o as inputs:

$$L(s_h, s_o) = \mathcal{P}(s_h) * \mathcal{P}(s_o), \tag{2}$$

where

$$\mathcal{P}(x) = \frac{T}{1 + e^{(k - wx)}}, \tag{3}$$

$\mathcal{P}(\cdot)$ is a part of the logistic function, the value of $\mathbf{T}$, $\mathbf{k}$ and $\mathbf{w}$ will be determined by data-driven manner. Figure 5 depicts the curve of $\mathcal{P}(\cdot)$ whose domain definition is $(0, 1)$. Bounding boxes will have low weight till their score is higher than a threshold. Previous works [31, 19] often directly multiply detection scores by the final classification score. But they cannot notably emphasize the differentiation between high quality and inaccurate detection results. LIS has the ability to enhance the differentiation between high and low grade object detections as shown in Figure 5.

Weights Sharing Strategy. An additional benefit of our interactiveness network is that, if cooperated with multi-stream HOI detection model $\mathbf{C}$, $\mathbf{P}$ can share the weights of convolutional blocks with the ones in $\mathbf{C}$. As shown in Figure 3, blocks $H^{\mathbf{P}}$ and $O^{\mathbf{P}}$ can share weights with H^{C} and O^{C} in the joint training. This weights sharing strategy can guarantee information sharing and better optimization of $\mathbf{P}$ and $\mathbf{C}$ in the multi-task training.

4.4. Interactiveness Knowledge Transfer Training

With $\mathbf{R}$, $\mathbf{P}$ and $\mathbf{C}$, our framework has two modes of utilization: hierarchical joint training in *Default Mode*, and interactiveness transfer training in *Transfer Learning Mode*.

Hierarchical Joint Training. In *Default Mode*, we introduce our hierarchical joint training scheme, as illustrated in Figure 6 (a). By adding a supervisor $\mathbf{P}$, our framework works in an unconventional training mode. To be specific, the framework is trained with hierarchical classification tasks, *i.e.* explicit interactiveness discrimination and HOI classification. The objective function of the framework can be expressed as:

$$\mathcal{L} = \mathcal{L}^{\mathbf{C}} + \mathcal{L}^{\mathbf{P}}, \tag{4}$$

where $\mathcal{L}^{\mathbf{C}}$ denotes the HOI classification cross entropy loss, while $\mathcal{L}^{\mathbf{P}}$ is the binary classification cross entropy loss.

Different from one-stage methods, additional interactiveness discrimination enforces the model to learn interactiveness knowledge, which can bring more powerful supervised constraints. Namely, when a pair is predicted as specific HOIs such as "cut cake", $\mathbf{P}$ must give the prediction "interactive" simultaneously. Experiment results (Section 5.4) prove that interactiveness knowledge learning can effectively refine the training and improve the performance. The framework in Default Mode is called "$\mathbf{RP}_D\mathbf{C}_D$" in the following, where "*D*" indicates "Default".

Interactiveness Knowledge Transfer Training. Noting that $\mathbf{P}$ only needs binary labels which are beyond the HOI classes, so interactiveness is transferable and reusable. In *Transfer Learning Mode*, $\mathbf{P}$ can be used as a transferable knowledge learner to learn interactiveness from multiple datasets and be applied to each of them respectively, as illustrated in Figure 6 (b). On the contrary, $\mathbf{C}$ must be trained on a single dataset once a time considering the variety of HOI category settings in different datasets. Therefore, knowledge of the specific HOIs is difficult to transfer. We will compare and evaluate the transferability of interactiveness knowledge and HOI knowledge in Section 5.

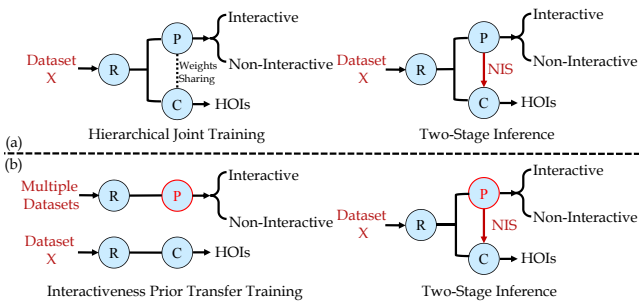

Figure 6. The schemes for training and testing. (a) In Default Mode, $\mathbf{P}$ and $\mathbf{C}$ are first trained jointly with weights sharing on the same dataset. (b) In Transfer Learning Mode, $\mathbf{P}$ can learn interactiveness knowledge across datasets and cooperates with multiple $\mathbf{C}$s trained on different datasets. In testing, our framework infers in two stages, *i.e.* $\mathbf{P}$ performs interactiveness discrimination at first, then $\mathbf{C}$ classifies the remaining edges/pairs.

For better representation of the transferability and performance enhancement of interactiveness, we set several transfer learning modes, referred as "$\mathbf{RP}_{Tn}\mathbf{C}_D$", where "*T*" indicates "Transfer", and "*n*" means $\mathbf{P}$ learns interactiveness knowledge from "*n*" datasets: **1)** $\mathbf{RP}_{T1}\mathbf{C}_D$: train $\mathbf{P}$ on 1 dataset and apply $\mathbf{P}$ to another dataset. **2)** $\mathbf{RP}_{T2}\mathbf{C}_D$: train $\mathbf{P}$ on 2 datasets and apply $\mathbf{P}$ to them respectively.

To compare the transferability of interactiveness knowledge and HOIs knowledge, we set a transfer learning mode "$\mathbf{RC}_T$" for $\mathbf{C}$: **3)** $\mathbf{RC}_T$**:** train $\mathbf{C}$ (without $\mathbf{P}$) on one dataset and apply it to another dataset. For example, we first train and test $\mathbf{C}$ on HICO-DET (referred as "$\mathbf{RC}_D$"). Second, we replace the last FC layer of $\mathbf{C}$ with a FC layer that fits the number of V-COCO HOIs, then finetune $\mathbf{C}$ for 1 epoch on V-COCO train set. Last, we test this new $\mathbf{C}$ on V-COCO test set. Details of the above modes can be found in Table 1.

4.5. Testing with Non-Interaction Suppression

After the interactiveness learning, we further utilize $\mathbf{P}$ to suppress the non-interactive pair candidates in testing, *i.e.* Non-Interaction Suppression (NIS). The inference process is based on tree structure as shown in Figure 2. Detected instances in test set will be paired exhaustively, so a dense graph $\mathcal{G}$ of human and objects is generated. First, we employ $\mathbf{P}$ to compute the interactiveness score of all edges. Next, we suppress the edges that meet NIS conditions, *i.e.* interaction score $s^{\mathbf{P}}_{(h,o)}$ smaller than a certain threshold α.

Through NIS, we can convert $\mathcal{G}$ to $\mathcal{G}^{'}$ where $\mathcal{G}^{'}$ denotes the approximate sparse HOI graph. The HOI classification score vector $\mathcal{S}^{\mathbf{C}}_{(h,o)}$ of (v_h, v_o) from $\mathbf{C}$ is:

$$\mathcal{S}^{\mathbf{C}}_{(h,o)} = \mathcal{F}_{\mathbf{C}}[\Gamma^{'}; \mathcal{G}^{'}(v_h, v_o)], \tag{5}$$

where $\Gamma^{'}$ are input features. The final HOI score vector of a pair (v_h, v_o) can be obtained by:

$$\mathcal{S}_{(h,o)} = \mathcal{S}^{\mathbf{C}}_{(h,o)} * s^{\mathbf{P}}_{(h,o)}. \tag{6}$$

Here we multiply interactiveness score $s^{\mathbf{P}}_{(h,o)}$ from $\mathbf{P}$ by the output of $\mathbf{C}$.

5. Experiments

In this section, we first introduce datasets and metrics adopted and then give implementation details of our framework. Next, we report our HOI detection results quantitatively and qualitatively compared with state-of-the-art approaches. Finally, we conduct ablation studies to validate the validity of the components in our framework.

5.1. Datasets and Metrics

Datasets. We adopt two HOI datasets HICO-DET [34] and V-COCO [13]. **HICO-DET [34]** includes 47,776 images (38,118 in train set and 9658 in test set), 600 HOI categories on 80 object categories (same with [7]) and 117 verbs, and provides more than 150k annotated human-object pairs. **V-COCO [13]** provides 10,346 images (2,533 for training, 2,867 for validating and 4,946 for testing) and 16,199 person instances. Each person has annotations for 29 action categories (five of them have no paired object). The objects are divided into two types: "object" and "instrument".

Metrics. We follow the settings adopted in [34], *i.e.* a prediction is a true positive only when the human and object bounding boxes both have IoUs larger than 0.5 with reference to ground truth, and the HOI classification result is accurate. The role mean average precision [13] is used to measure the performance.

5.2. Implementation Details

We employ a Faster R-CNN [11] with ResNet-50 [14] as $\mathbf{R}$ and keep it frozen. $\mathbf{C}$ consists of three streams similar to [34, 31], extracting features $\mathbf{\Gamma}^{'}$ from instance appearance, spatial location as well as context. Within human and object streams, a residual block [14] with global average pooling and four 1024 sized FCs are used. Relatively, the spatial stream is composed of two convolutional layers with max pooling, and two 1024 sized FCs. Following [34, 31], we use the late fusion strategy in $\mathbf{C}$. $\mathbf{P}$ also consists of three streams (seen in Figure 3). A residual block [14] with global average pooling, and two 1024 sized FCs are adopted in human and object streams. Residual blocks within these two streams will share weights with those in $\mathbf{C}$. Spatial-Pose stream consists of two convolutional layers with max pooling and two 1024 sized FCs. The outputs of three streams are concatenated and passed through two 1024 sized FCs to perform interactiveness discrimination.

Test Set	Method	$\mathbf{P}$-Train Set	$\mathbf{C}$-Train Set
HICO-DET	$\mathbf{RP}_D\mathbf{C}_D$	HICO-DET	HICO-DET
	$\mathbf{RP}_{T1}\mathbf{C}_D$	V-COCO	HICO-DET
	$\mathbf{RP}_{T2}\mathbf{C}_D$	HICO-DET, V-COCO	HICO-DET
HICO-DET	$\mathbf{RC}_D$	-	HICO-DET
	$\mathbf{RC}_T$	-	V-COCO
V-COCO	$\mathbf{RP}_D\mathbf{C}_D$	V-COCO	V-COCO
	$\mathbf{RP}_{T1}\mathbf{C}_D$	HICO-DET	V-COCO
	$\mathbf{RP}_{T2}\mathbf{C}_D$	HICO-DET, V-COCO	V-COCO
V-COCO	$\mathbf{RC}_D$	-	V-COCO
	$\mathbf{RC}_T$	-	HICO-DET

Table 1. Mode settings in experiments.

For a fair comparison, we adopt the object detection results and COCO [7] pre-trained weights from [31] which are provided by authors. Since NIS and LIS can suppress non-interactive pairs, we set detection confidence thresholds lower than [31], *i.e.* 0.6 for human and 0.4 for object. The image-centric training strategy [11] is also applied. In other words, pair candidates from one image make up the mini-batch. We adopt SGD and set an initial learning rate as 1e-4, weight decay as 1e-4, momentum as 0.9. In training, the ratio of positive and negative samples is 1:3. We jointly train the framework for 25 epochs. In LIS mentioned in Equation 3, we set $T = 8.4, k = 12.0, w = 10.0$. In testing, the interactiveness threshold α in NIS is set as 0.1. All experiments are conducted on a single Nvidia Titan X GPU.

5.3. Results and Comparisons

We compare our method with five state-of-the-art HOI detection methods [34, 37, 19, 32, 31] on HICO-DET, and four methods [13, 19, 32, 31] on V-COCO. The HOI detection result is evaluated with mean average precision. For HICO-DET, we follow the settings in [34]: Full (600 HOIs), Rare (138 HOIs), Non-Rare (462 HOIs) in Default and Known Object mode. For V-COCO, we evaluate AP_{role} (24 actions with roles). More details can be found in [34, 13].

Default Mode. From Table 2, we can find that the $\mathbf{RP}_D\mathbf{C}_D$ has already outperformed compared methods. We respectively achieve **17.03** and **19.17** mAP on Default and Know Object Full sets on HICO-DET. In particular, we boost the performance of **2.97** and **4.18** mAP on Rare sets. To illustrate, as the generalization ability of interactiveness is beyond HOI category settings, information scarcity and learning difficulty of rare categories is alleviated. So the performance difference between rare and non-rare categories is accordingly reduced. Results on V-COCO are shown in Table 3. $\mathbf{RP}_D\mathbf{C}_D$ also achieves superior performance and outperforms state-of-the-art method [31] (late and early fusion model), yielding **47.8** mAP, which quan-

Method	Default			Known Object		
	Full	Rare	Non-Rare	Full	Rare	Non-Rare
Shen *et al.* [37]	6.46	4.24	7.12	-	-	-
HO-RCNN [34]	7.81	5.37	8.54	10.41	8.94	10.85
InteractNet [19]	9.94	7.16	10.77	-	-	-
GPNN [32]	13.11	9.34	14.23	-	-	-
iCAN [31]	14.84	10.45	16.15	16.26	11.33	17.73
$\mathbf{RC}_D$	13.75	10.23	15.45	15.34	10.98	17.02
$\mathbf{RP}_D\mathbf{C}_D$	**17.03**	**13.42**	**18.11**	**19.17**	**15.51**	**20.26**
$\mathbf{RC}_T$	10.61	7.78	11.45	12.47	8.87	13.54
$\mathbf{RP}_{T1}\mathbf{C}_D$	**16.91**	**13.32**	**17.99**	**19.05**	**15.22**	**20.19**
$\mathbf{RP}_{T2}\mathbf{C}_D$	**17.22**	**13.51**	**18.32**	**19.38**	**15.38**	**20.57**

Table 2. Results comparison on HICO-DET [34]. D indicates the default mode, and T means the transfer learning model.

titatively validates the efficacy of the interactiveness. Notably, $\mathbf{RC}_D$ shows limited performance when compared with other models containing $\mathbf{P}$. This reveals the performance enhancement ability of interactiveness network $\mathbf{P}$.

Transfer Learning Mode. By leveraging transferred interactiveness knowledge, $\mathbf{RP}_{T2}\mathbf{C}_D$ presents great performance improvement and achieves the most state-of-the-art performance. On HICO-DET, $\mathbf{RP}_{T2}\mathbf{C}_D$ surpasses [31] by **2.38**, **3.06**, and **2.17** mAP on three Default category sets. Meanwhile, it also outperforms [31] by **4.0** and **3.4** mAP on V-COCO. This indicates the good transferability and effectiveness of interactiveness. Since HICO-DET train set (38K) is much bigger than V-COCO train set (2.5K), improvement is also larger when transferring is performed from HICO-DET to V-COCO. As we can see, mode $\mathbf{RP}_{T1}\mathbf{C}_D$ achieves obvious improvement on V-COCO, but it shows a relatively smaller improvement on HICO-DET when compared with mode $\mathbf{RP}_D\mathbf{C}_D$.

We also evaluate the transferability of HOIs knowledge. In comparison with $\mathbf{RC}_D$, $\mathbf{RC}_T$ shows a significant performance decrease of **3.14** and **4.7** mAP on two datasets, as shown in Table 2 and 3. It proves that interactiveness is more suitable and easier to transfer than HOIs knowledge.

Non-Interaction Reduction. The non-interactive pairs reduction effect after employing NIS are shown in Table 4. In default mode $\mathbf{RP}_D\mathbf{C}_D$, NIS shows obvious effectiveness. With interactiveness transferred from multiple datasets, $\mathbf{RP}_{T2}\mathbf{C}_D$ achieves better suppressive effect and discards **70.94%** and **73.62%** non-interactive pairs respectively on two datasets, thus bringing more performance gain. Meanwhile, $\mathbf{RP}_{T1}\mathbf{C}_D$ also performs well and suppresses a certain amount of non-interactive pair candidates. This suggests the good transferability of interactiveness.

Visualized Results. Representative predictions are shown in Figure 7. We can find that our model is capable of detecting various kinds of complicated HOIs such as multiple interactions within one pair, one person performing multiple interactions with different objects, one object interacted with multiple persons, multiple persons performing different interactions with multiple objects.

Figure 8 shows the visualized effects of NIS. We can see that NIS effectively distinguish the non-interactive pairs and suppress them in extremely difficult scenarios, such as a person performing a confusing action and the tennis ball, a crowd of people with ties. In the bottom-left corner we show an even harder sample. When the subject and object are the left hand and right hand, $\mathbf{C}$ predicts wrong HOI "type_on keyboard". $\mathbf{C}$ may mistake the left hand for the keyboard because they are too close. However, $\mathbf{P}$ accurately figures out that two hands are non-interactive. These results prove that the one-stage method would yield many false positives without interactiveness and NIS.

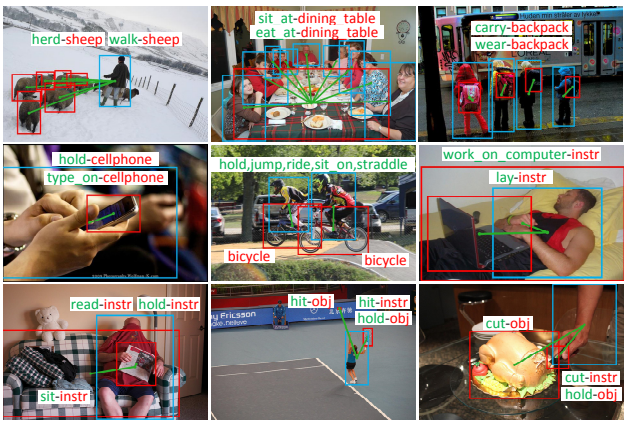

Figure 7. Visualization of sample HOI detections. Subjects and objects are represented with blue and red bounding boxes. While interactions are marked by green lines linking the box centers.

Method	AP_{role}
Gupta *et al.* [13]	31.8
InteractNet [19]	40.0
GPNN [32]	44.0
iCAN w/ late(early) [31]	44.7 (45.3)
$\mathbf{RC}_D$	43.2
$\mathbf{RP}_D\mathbf{C}_D$	**47.8**
$\mathbf{RC}_T$	38.5
$\mathbf{RP}_{T1}\mathbf{C}_D$	**48.3**
$\mathbf{RP}_{T2}\mathbf{C}_D$	**48.7**

Table 3. Results comparison on V-COCO [13]. D indicates the default mode, and T means the transfer learning model.

5.4. Ablation Studies

In mode $\mathbf{RP}_D\mathbf{C}_D$, we analyze the significance of Low-grade Instance Suppressive, Non-Interaction Suppression and the three streams within $\mathbf{P}$ (seen in Table 5).

Non-Interaction Suppression NIS plays a key role to reduce the non-interactive pairs. We evaluate its impact by removing NIS during testing. In other words, we directly use the $\mathcal{S}_{(h,o)}$ from Equation 6 as the final predictions without NIS. Consequently, the model shows an obvious performance degradation, which proves the importance of NIS.

Low-grade Instance Suppressive LIS suppress the low-grade object detections and reward the high-grade ones. By removing $L(s_h, s_o)$ in Equation 1, we observe a degradation in Table 5. This suggests that LIS is capable of distinguishing the low-grade detections and improves the performance

Figure 8. Visualized effects of NIS. Green lines mean accurate HOIs, while purple lines mean non-interactive pairs which are suppressed. Without NIS, **C** would generate false positive predictions for these non-interactive pairs in one-stage inference, which are shown by the purple texts below the images. Even some extremely hard scenarios can be discovered and suppressed, such as mis-groupings between person and object close to each other, person and object in clutter scene.

Test Set	Method	Reduction
HICO-DET	$\mathbf{RP}_D\mathbf{C}_D$	-65.96%
	$\mathbf{RP}_{T1}\mathbf{C}_D$	-62.24%
	$\mathbf{RP}_{T2}\mathbf{C}_D$	-70.94%
V-COCO	$\mathbf{RP}_D\mathbf{C}_D$	-65.98%
	$\mathbf{RP}_{T1}\mathbf{C}_D$	-59.51%
	$\mathbf{RP}_{T2}\mathbf{C}_D$	-73.62%

Table 4. Non-interactive pairs reduction after performing NIS.

	HICO-DET		V-COCO
Method	Default Full	KO Full	AP_{role}
$\mathbf{RP}_D\mathbf{C}_D$	17.03	19.17	47.8
w/o NIS	15.86	17.35	46.2
w/o LIS	16.35	18.83	47.4
w/o NIS & LIS	15.45	17.31	45.8
H Stream Only	14.91	16.21	44.5
O Stream Only	15.28	16.89	45.2
S-P Stream Only	15.73	17.46	46.0

Table 5. Results of ablation studies. Human, object, spatial-pose stream are representated as H, O and S-P stream.

without using more costly superior object detector.

NIS & LIS Without NIS and LIS both, our method only takes effect in the joint training of **P** and **C**. As we can see in Table 5, performance degrades greatly but still outperforms other methods, which indicates the enhancement brought by **P** in the hierarchical joint training.

Three Streams. By keeping one stream in **P** each time, we evaluate their contributions as shown in Table 5. We can find that spatial-pose stream is the largest contributor, but we still need appearance features from the other two streams to achieve better performance.

6. Conclusion

In this paper, we propose a novel method to learn and utilize the implicit interactiveness knowledge, which is general and beyond HOI categories. Thus, it can be transferred across datasets. With interactiveness knowledge, we exploit an interactiveness network to perform Non-interaction Suppression before HOI classification in inference. Extensive experiment results show the efficacy of interactiveness. By combining our method with existing detection models, we achieve state-of-the-art results on HOI detection.

Acknowledgement: This work is supported in part by the National Key R&D Program of China, No.2017YFA0700800, National Natural Science Foundation of China under Grants 61772332.

References

[1] Yang Wang, Hao Jiang, Mark S Drew, Ze-Nian Li, and Greg Mori. Unsupervised discovery of action classes. In *CVPR*, 2006. 2

[2] N Ikizler, R. G Cinbis, S Pehlivan, and P Duygulu. Recognizing actions from still images. In *ICPR*, 2008. 2

[3] Brenna D Argall, Sonia Chernova, Manuela Veloso, and Brett Browning. A survey of robot learning from demonstration. *Robotics and autonomous systems*, 57(5), 2009. 1

[4] Weilong Yang, Yang Wang, and Greg Mori. Recognizing human actions from still images with latent poses. In *CVPR*, 2010. 2

[5] Vincent Delaitre, Ivan Laptev, and Josef Sivic. Recognizing human actions in still images: a study of bag-of-features and part-based representations. In *BMVC*, 2010. 2

[6] M. A. Sadeghi and A. Farhadi. Recognition using visual phrases. In *CVPR*, 2012. 2

[7] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 4, 6

[8] Chao-Yeh Chen and Kristen Grauman. Predicting the location of interactees in novel human-object interactions. In *ACCV*, 2014. 2

[9] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *CVPR*, 2015. 1

[10] Ross Girshick. Fast r-cnn. In *ICCV*, 2015. 3

[11] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*, 2015. 1, 3, 6

[12] Yu Wei Chao, Zhan Wang, Yugeng He, Jiaxuan Wang, and Jia Deng. Hico: A benchmark for recognizing human-object interactions in images. In *ICCV*, 2015. 2

[13] Saurabh Gupta and Jitendra Malik. Visual semantic role labeling. *arXiv preprint arXiv:1505.04474*, 2015. 1, 2, 6, 7

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 3, 6

[15] Arun Mallya and Svetlana Lazebnik. Learning models for actions and person-object interactions with transfer to question answering. In *ECCV*, 2016. 2

[16] Mark Yatskar, Luke Zettlemoyer, and Ali Farhadi. Situation recognition: Visual semantic role labeling for image understanding. In *CVPR*, 2016. 2

[17] Cewu Lu, Ranjay Krishna, Michael Bernstein, and Li Fei-Fei. Visual relationship detection with language priors. In *ECCV*, 2016. 1, 2

[18] Bohan Zhuang, Qi Wu, Chunhua Shen, Ian Reid, and Anton van den Hengel. Care about you: towards large-scale human-centric visual relationship detection. *arXiv preprint arXiv:1705.09892*, 2017. 2

[19] Georgia Gkioxari, Ross Girshick, Piotr Dollár, and Kaiming He. Detecting and recognizing human-object interactions. *arXiv preprint arXiv:1704.07333*, 2017. 1, 2, 3, 5, 6, 7

[20] Tsung-Yi Lin, Piotr Dollár, Ross B Girshick, Kaiming He, Bharath Hariharan, and Serge J Belongie. Feature pyramid networks for object detection. In *CVPR*, 2017. 3

[21] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, 2017. 1

[22] Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu. RMPE: Regional multi-person pose estimation. In *ICCV*, 2017. 4

[23] Danfei Xu, Yuke Zhu, Christopher B Choy, and Li Fei-Fei. Scene graph generation by iterative message passing. In *CVPR*, 2017. 2

[24] 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. *IJCV*, 123(1):32–73, 2017. 1, 2

[25] Hanwang Zhang, Zawlin Kyaw, Shih-Fu Chang, and Tat-Seng Chua. Visual translation embedding network for visual relation detection. In *CVPR*, 2017. 2

[26] Cewu Lu, Hao Su, Yonglu Li, Yongyi Lu, Li Yi, Chi-Keung Tang, and Leonidas J Guibas. Beyond holistic object recognition: Enriching image understanding with part states. In *CVPR*, 2018. 1

[27] Jiefeng Li, Can Wang, Hao Zhu, Yihuan Mao, Hao-Shu Fang, and Cewu Lu. Crowdpose: Efficient crowded scenes pose estimation and a new benchmark. *arXiv preprint arXiv:1812.00324*, 2018. 4

[28] Bo Pang, Kaiwen Zha, Hanwen Cao, Chen Shi, and Cewu Lu. Deep rnn framework for visual sequential applications. *arXiv preprint arXiv:1811.09961*, 2018. 1

[29] Ross Girshick, Ilija Radosavovic, Georgia Gkioxari, Piotr Dollár, and Kaiming He. Detectron. https://github.com/facebookresearch/detectron, 2018. 3

[30] Jianwei Yang, Jiasen Lu, Stefan Lee, Dhruv Batra, and Devi Parikh. Graph r-cnn for scene graph generation. In *ECCV*, 2018. 2

[31] Chen Gao, Yuliang Zou, and Jia-Bin Huang. ican: Instance-centric attention network for human-object interaction detection. *arXiv preprint arXiv:1808.10437*, 2018. 1, 2, 3, 5, 6, 7

[32] Siyuan Qi, Wenguan Wang, Baoxiong Jia, Jianbing Shen, and Song-Chun Zhu. Learning human-object interactions by graph parsing neural networks. In *ECCV*, 2018. 1, 2, 6, 7

[33] Haoshu Fang, Yuanlu Xu, Wenguan Wang, Xiaobai Liu, and Song-Chun Zhu. Learning pose grammar to encode human body configuration for 3d pose estimation. In *AAAI*, 2018. 1

[34] Yu-Wei Chao, Yunfan Liu, Xieyang Liu, Huayi Zeng, and Jia Deng. Learning to detect human-object interactions. In *WACV*, 2018. 1, 2, 3, 4, 6, 7

[35] Hao-Shu Fang, Jinkun Cao, Yu-Wing Tai, and Cewu Lu. Pairwise body-part attention for recognizing human-object interactions. In *ECCV*, 2018. 2

[36] Yuliang Xiu, Jiefeng Li, Haoyu Wang, Yinghong Fang, and Cewu Lu. Pose flow: Efficient online pose tracking. *arXiv preprint arXiv:1802.00977*, 2018. 1

[37] Liyue Shen, Serena Yeung, Judy Hoffman, Greg Mori, and Fei Fei Li. Scaling human-object interaction recognition through zero-shot learning. In *WACV*, 2018. 2, 6, 7

[38] Wenqiang Xu, Yonglu Li, and Cewu Lu. Srda: Generating instance segmentation annotation via scanning, reasoning and domain adaptation. In *ECCV*, 2018. 1

[39] Hao-Shu Fang, Guansong Lu, Xiaolin Fang, Jianwen Xie, Yu-Wing Tai, and Cewu Lu. Weakly and semi supervised human body part parsing via pose-guided knowledge transfer. *CVPR*, 2018. 1

[40] Guojun Yin, Lu Sheng, Bin Liu, Nenghai Yu, Xiaogang Wang, Jing Shao, and Chen Change Loy. Zoom-net: Mining deep feature interactions for visual relationship recognition. *arXiv preprint arXiv:1807.04979*, 2018. 2