This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2021.3054048, IEEE Transactions on Pattern Analysis and Machine Intelligence

Transferable Interactiveness Knowledge for Human-Object Interaction Detection

Yong-Lu Li, Xinpeng Liu, Xiaoqian Wu, Xijie Huang, Liang Xu, Cewu Lu, *Member, IEEE*

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 and alleviate the gap between diverse 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 utilize the human instance and body part features together to learn the interactiveness in hierarchical paradigm, i.e., instance-level and body part-level interactivenesses. Thereafter, a consistency task is proposed to guide the learning and extract deeper interactive visual clues. We extensively evaluate the proposed method on HICO-DET, V-COCO, and a newly constructed HAKE-HOI dataset. With the learned interactiveness, our method outperforms state-of-the-art HOI detection methods, verifying its efficacy and flexibility. Code is available at https://github.com/DirtyHarryLYL/Transferable-Interactiveness-Network.

Index Terms—Human-Object Interaction, Interactiveness, Transfer Learning.

1 Introduction

Human-Object Interaction (HOI) detection retrieves human and object locations and infers the interaction classes simultaneously from still image. As a sub-task of visual relationship [1], [2], HOI is strongly related to the human body and object understanding [3], [4], [5], [6]. It is crucial for behavior understanding and can facilitate activity understanding [7], imitation learning [8], etc. Recently, impressive progress has been made by utilizing Deep Neural Networks (DNNs) in this area [9], [10], [11], [12].

Generally, human and objects need to be detected first in HOI detection. Given an image and its detections, human and objects are often paired exhaustively [10], [11], [12]. HOI detection task aims to classify these pairs as various HOI categories. Previous one-stage methods [9], [10], [11], [12], [13] 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 "eat apple", we can implicitly predict that it is interactive.

Though interactiveness is an essential element for HOI detection, previous methods neglected to study how to utilize it and improve its learning. In comparison to various HOI categories, interactiveness conveys more basic information. Such an attribute makes it easier to transfer across datasets. Based on this inspiration, we propose a **Interactiveness Knowledge** learning method as seen in Fig. 1. With our method, interactiveness can be learned across datasets and applied to any specific datasets. By utilizing interactiveness, we take two stages to identify HOIs: first discriminate a human-object pair as interactive or not and then classify it as specific HOIs. Compared to previous one-stage method [9], [10], [11], [12], [13], 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 propose a novel two-stage method to classify pairs hierarchically as shown in Fig. 2. Our model, Transferable Interactiveness Network (TIN), con-

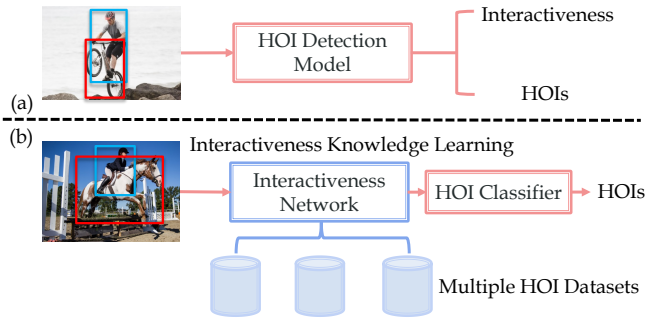

Fig. 1: Interactiveness Knowledge Learning. (a) HOI datasets contain implicit interactiveness knowledge. We can learn it better by performing **explicit** interactiveness discrimination, and utilize it to improve the HOI detection performance. (b) Interactiveness knowledge is beyond the HOI categories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

- *Yong-Lu Li, Xinpeng Liu, Xiaoqian Wu, Xijie Huang, Liang Xu and Cewu Lu are with the Department of Electrical and Computer Engineering, Shanghai Jiao Tong University, Shanghai, 200240, China. E-mail: {yonglu_li, enlighten, otaku_huang, liangxu}@sjtu.edu.cn, xinpengliu0907@gmail.com.*
- *Cewu Lu is the corresponding author, member of Qing Yuan Research Institute and MoE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University, China and Shanghai Qi Zhi institute. E-mail: lucewu@sjtu.edu.cn.*

0162-8828 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Carleton University. Downloaded on May 26,2021 at 15:21:11 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2021.3054048, IEEE Transactions on Pattern Analysis and Machine Intelligence

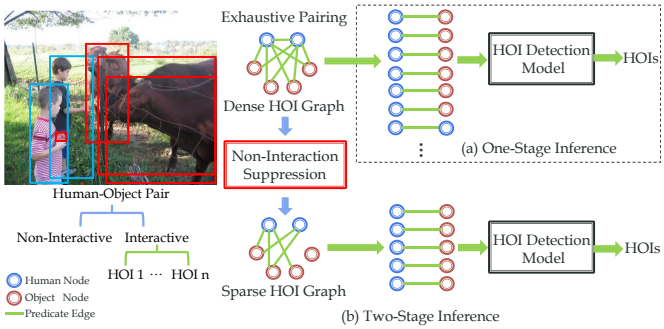

Fig. 2: HOIs within an image can be represented as a HOI graph. Human and object can be seen as nodes, whilst the interactions are represented as edges. Exhaustive pairing of all nodes would import overmuch non-interactive edges and do damage to detection performance. Our Non-Interaction Suppression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

sists of three networks: Representation Network (extractor, referred as $\mathbf{R}$), HOI Network (classifier, referred as $\mathbf{C}$) and Interactiveness Network (discriminator, referred as $\mathbf{D}$). The interactiveness network $\mathbf{D}$ is creatively utilized for binary classification, i.e., interactiveness/non-interactiveness. It benefits the whole model in two aspects.

For one thing, the conventional HOI model is only targeted at HOI detection and classification. Our HOI classifier $\mathbf{C}$ can be trained together with the interactiveness discriminator $\mathbf{D}$ to learn the HOIs and interactiveness knowledge together. Under usual circumstances, the ratio of non-interactive edges is dominant within inputs. Thus, by utilizing binary interactiveness labels converted from HOI labels, the whole model would be trained with a stronger supervised constraint and performs better and more robustly.

For another, noting that interactiveness network $\mathbf{D}$ only needs binary labels which are beyond the HOI classes, interactiveness is **transferable** and **reusable**. Therefore, $\mathbf{D}$ can be used as a transferable knowledge learner to learn interactiveness from multiple datasets and be applied to each of them respectively.

In testing, we adopt the two-stage policy. First, the interactiveness network $\mathbf{D}$ evaluates the interactiveness of a human-object pair (edge) by exploiting the learned interactiveness knowledge, so we can convert the dense HOI graph to a sparse one (Fig. 2). After this, $\mathbf{C}$ will process the sparse graph and classify the remaining edges.

To implement TIN, we propose a hierarchical framework. First, we utilize the human/object appearance and spatial configuration as the instance-level features to learn the *interactiveness between instances*. Second, we further argue that interactiveness has an important characteristic related to human body parts. That is, when interacting with daily objects, only some parts of our body would get involved. For example, in "read book", only our head and hands have strong relationships with the book, but not our lower body. We can either stand or lie while reading. In view of this, besides the de facto instance-level features, we further define the interactiveness between object and human body parts, i.e., *part interactiveness*. Then, human body part feature paired with object feature are used to learn it. Notably, instance and part interactivenesses have inherent and implicit relationships. Their relationship is in line with the Multi-Instance Learning (MIL) [14], i.e., the instance interactiveness is *false* if and only if all part interactivenesses are *false*. To be more explicit, a human is interacting with an object if and only if at least one human part is interacting with the object. Thus, when inputting different level features, we can construct this consistency between two levels as objective in learning. Moreover, body parts with higher interactiveness score should be paid more attention. We further use part attention strategy to strengthen the important parts in HOI inference. The experiment (Sec. 5.5.3) verifies our assumption that different HOIs have various part interactiveness patterns. For instance, "ride" is learned to be more related to feet, thighs and hands than head and hip. Thus, such attention policy can greatly benefits the HOI learning.

We perform extensive experiments on HICO-DET [9], V-COCO [13] and a newly constructed dataset HAKE-HOI [15]. Our method cooperated with transferred interactiveness outperforms the state-of-the-art methods by **1.53** and **4.35** mAP on the Default set and Rare set of HICO-DET.

2 Related Works

Visual Relationship Detection. Visual relationship detection [1], [2], [16], [17] aims to detect the objects and classify their relationships simultaneously. In [2], 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 [1] is then proposed to promote the studies in this direction. Recent works [18], [19], [20], [21] 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 [22], [23], [24] is essential to understand human-centric interaction with objects. Several large-scale datasets, such as V-COCO [13], HICO-DET [9], HCVRD [25], HAKE [15] were proposed for the exploration of HOI detection. Different from HOI recognition [26], [27], [28], [29], [30] 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, recent methods have made significant progress.

Chao et al. [9] proposed a multi-stream model combining visual features, spatial locations to help tackle this problem. To address the long-tail issue, Shen et al. [31] studied zero-shot learning and predicted the verb and object separately. In InteractNet [10], an action-specific density map estimation method is introduced to locate interacted objects. In [11], Qi et al. proposed GPNN incorporating DNN and graph model, which uses message parsing to iteratively update states and classifies all possible pairs/edges. Gao et al. [12] exploited an instance-centric attention module to enhance the information from interest regions and facilitate the HOI classification. Peyre et al. [32] learned a unified space combining visual and semantic language features and detected unseen interactions through entity analogy.

0162-8828 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Carleton University. Downloaded on May 26,2021 at 15:21:11 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2021.3054048, IEEE Transactions on Pattern Analysis and Machine Intelligence

Generally, these methods infer HOI 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.

Part-based Action Recognition. Part-level human feature takes further insight into Human-Object Interaction. Based on the whole person and part bounding boxes, Gkioxari et al. [33] developed a part-based model to make fine-grained action recognition. In [26], Fang et al. proposed a new pairwise body part attention model that can learn to focus on crucial parts and their correlations for HOI recognition. A novel attention-based feature selection method and a pairwise parts representation learning scheme are introduced. In our work, we utilize the body part features and whole body feature to learn hierarchical interactivenesses. And the unique consistency between two levels is fully explored to guide the learning.

3 Preliminary

HOI representation can be described as a graph model [11], [18] as seen in Fig. 2. Instances and relations are expressed as nodes and edges respectively. With exhaustive pairing [10], [12], 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 HOI labeling on its edges.

4 Our Method

4.1 Overview

In this section, 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.

The overview of our TIN framework is shown in Fig. 3. We propose interactiveness network $\mathbf{D}$ (interactiveness discriminator) which utilizes interactiveness to reduce false positives caused by overmuch non-interactive pair candidates. 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 classify HOIs. In testing, $\mathbf{D}$ is utilized in two stages. First, $\mathbf{D}$ 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 $\mathbf{D}$, $\mathbf{C}$ will process the sparse graph and classify the remaining edges.

In subsequent sections, we first introduce the conventional modules $\mathbf{R}$ and $\mathbf{C}$ in Sec. 4.2. Then, the structure of interactiveness network $\mathbf{D}$ is detailed in Sec. 4.3. Finally, the Non-Interaction Suppression (NIS) is discussed in Sec. 4.4.

4.2 Representation and Classification Networks

First, we make a brief introduction to the representation network $\mathbf{R}$ and classification network $\mathbf{C}$.

Human and Object Detection. In HOI detection, human and objects need to be detected first. In this work, we follow the setting of [12] and employ the Detectron [34] with ResNet-50-FPN [35] 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 [9], [10], [12], $\mathbf{R}$ is often modified from object detector such as Fast R-CNN [36] or Faster R-CNN [4]. We also exploited a Faster R-CNN [4] with ResNet-50 [37] 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.

Classification Network. As for $\mathbf{C}$, multi-stream architecture and late fusion strategy are frequently used and proved effective [9], [12]. Follow [9], [12], 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 [37] (denoted as $H^{\mathbf{C}}$ and $O^{\mathbf{C}}$) with pooling layer and fully connected layers (FCs) are adopted. Moreover, an extra spatial stream [9] 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 Fig. 5. 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. The prediction scores of human and object streams will be fused by element-wise sum in the same proportion, then we multiply it with the score of spatial stream and produce the final result of $\mathbf{C}$.

4.3 Interactiveness Network

The interactiveness network $\mathbf{D}$ is designed for binary classification: interactive/non-interactive. As aforementioned, we infer the interactiveness in a hierarchical way: 1) **Instance-Level.** The human/object appearance and spatial configuration are used as the instance-level features to predict the interactiveness between human and object. 2) **Part-Level.** Notably, we further utilize the human body part features to take deeper insight into the interactivenesses between body parts and object. With interactivenesses from two levels, we then can use the consistency between them to guide the learning. To summarize, there are four kinds of streams (human, object, spatial-pose and part) in $\mathbf{D}$. Each of them focuses on different elements of the HOIs in images. The architecture of interactiveness network $\mathbf{D}$ is shown in Fig. 4.

0162-8828 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Carleton University. Downloaded on May 26,2021 at 15:21:11 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2021.3054048, IEEE Transactions on Pattern Analysis and Machine Intelligence

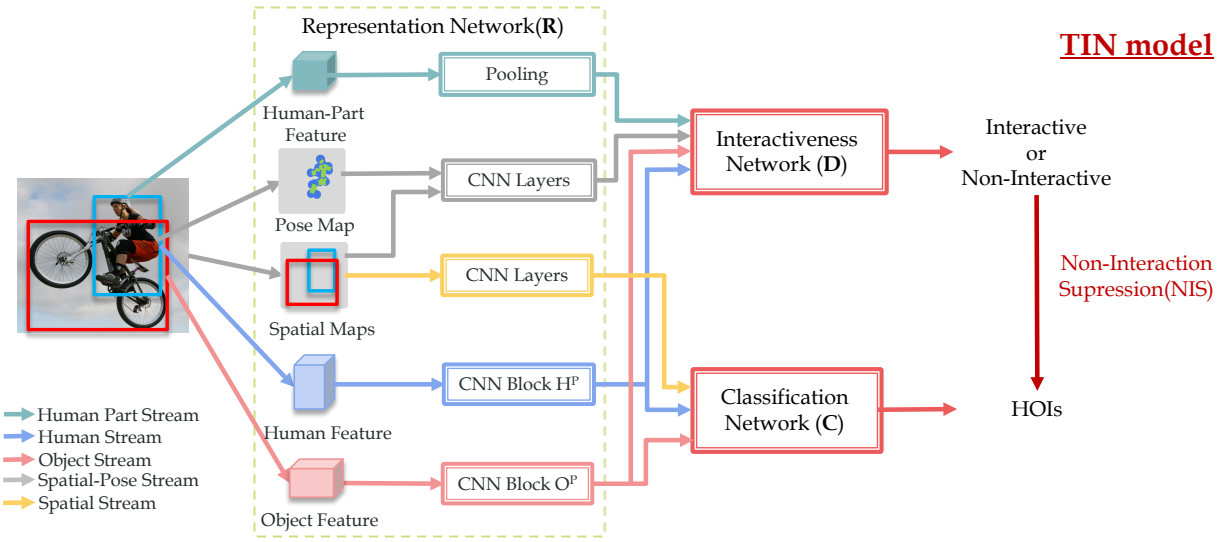

Fig. 3: The overview of our TIN framework. Interactiveness Network **D** utilizes interactiveness to reduce false positives caused by overmuch non-interactive pair candidates. Some conventional modules are also included, namely, Representation Network **R** and Classification Network **C**. **R** is responsible for feature extraction from detected instances. **C** utilizes node and edge features to perform HOI classification. In testing, **D** is utilized in two stages. First, **D** evaluates the interactiveness of edges by exploiting the learned interactiveness knowledge and impose NIS on **C**. Second, combined with interactiveness score from **D**, **C** will process the sparse graph and classify the remaining edges.

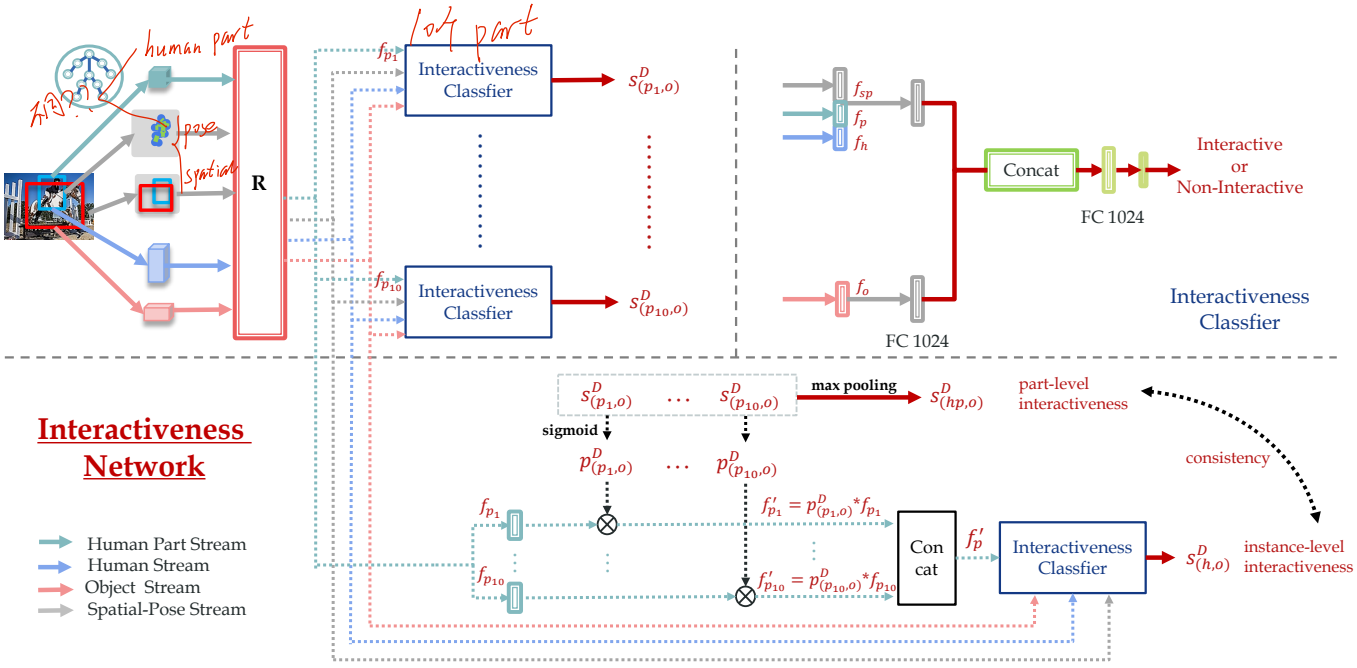

Fig. 4: The architeture of Interactiveness Network **D**. There are eleven interactiveness binary classifiers in **D**, i.e., ten for part interactivenesses and one for instance interactiveness (illustrated in the upper right part). For the part-level classifier, the i-th part feature f_{p_i} together with f_h, f_o and f_{sp}, are concatenated and input to FCs and Sigmoid to generate the part interactiveness probability $p^{\mathbf{D}}_{(p_i,o)}$. Meanwhile, we utilize part interactivenesses to select the important parts, i.e. $f'_{p_i} = p^{\mathbf{D}}_{(p_i,o)} * f_{p_i}$. For the instance-level classifier, we concatenate the ten re-weighted part features f'_{p_i} $(1 \leq i \leq 10)$ and input them to FCs to generate the instance interactivenss score $s^{\mathbf{D}}_{(h,o)}$. Finally, consistency between two levels is constructed as the objective.

In subsequent sections, we illustrate the details of the interactiveness network. First, we introduce three streams based on instance-level features in Sec. 4.3.1. Second, we explore the part stream based on part-level features in Sec. 4.3.2. Third, we detail the interactiveness binary classification via four streams in Sec. 4.3.3. Finally, consistency between two interactive levels is discussed in Sec. 4.3.4.

4.3.1 *Three Streams with Instance-Level Features*

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 Fig. 5, Person 1 and

0162-8828 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Carleton University. Downloaded on May 26,2021 at 15:21:11 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2021.3054048, IEEE Transactions on Pattern Analysis and Machine Intelligence

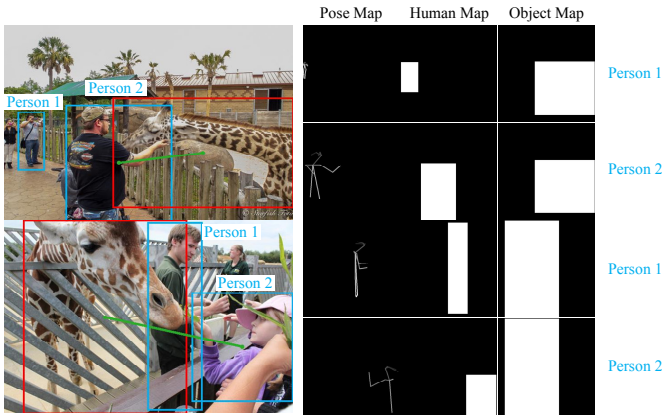

Fig. 5: Inputs of the spatial-pose stream. Three kinds of maps are included: pose map, human map and object map. Person 2 in two images both have interaction "feed" with giraffes. But two pairs of Person 1 and giraffe are all non-interactive. Their poses and locations are helpful for the interactiveness discrimination.

the giraffe far from him are not interactive. Their spatial maps [9] 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".

In view of this, we argue that the combination of visual appearance, spatial location and human pose information is the key to interactiveness discrimination. Hence $\mathbf{D}$ needs to encode all these key elements together to learn the interactiveness knowledge. A natural choice is a multi-stream architecture as presented: human, object and spatial-pose streams. Following [12], the instance-centric attention module is also adopted.

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 $H^{\mathbf{D}}$ and $O^{\mathbf{D}}$. The architecture of $H^{\mathbf{D}}$ and $O^{\mathbf{D}}$ is the same as $H^{\mathbf{C}}$ and $O^{\mathbf{C}}$ (Sec. 4.2). Through subsequent global average pooling and FCs, the output features of object stream is denoted as f_h and f_o.

Spatial-Pose Stream Different from [9], our spatial-pose stream input includes a special 64x64 pose map. Given the union box of each human and the paired object, we employ pose estimation [38] to estimate 17 body keypoints (in COCO format [39]). Then, we link the keypoints with lines of different gray values 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 other streams for the next interactiveness discrimination.

4.3.2 Part Stream with Part-Level Features

Conventional methods usually focus on instance interaction recognition, while body part-object interaction is often overlooked. This is partially caused by the difficulty to annotate the massive relationships between body parts and objects. However, such a relationship can be more easily defined via interactiveness. Therefore, we explore to adopt the human body part features in the interactiveness learning.

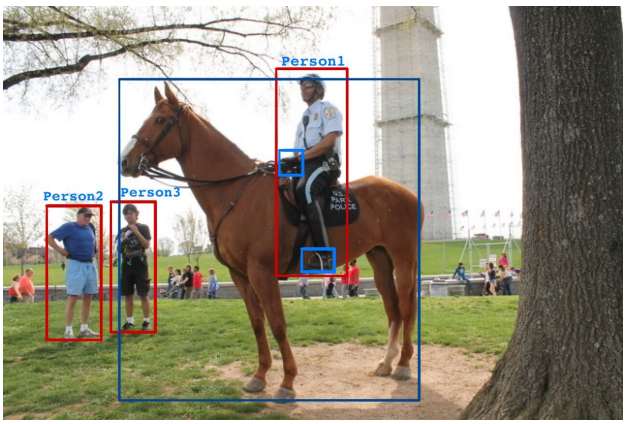

Fig. 6: The illustration of the relationship between part-level and instance-level interactiveness. Person 1 is inferred to interact with the horse from the features of his hands and feet, while Person 2 and Person 3 are not interacting with the horse based on the same reason.

To locate the parts, we first use the pose estimation [38] to construct ten part boxes (Fig. 6) following [26], i.e., head, upper arms, hands, hip, thighs and feet. Each part box is centered with a corresponding detected joint. And the size of a part box is decided by scaling the distance between the neck and pelvis joints. Second, for part stream, we extract the ROI pooling feature from the detected part box as the part feature, i.e., f_{p_i} for the i-th part, $1 \leq i \leq 10$.

4.3.3 Binary Interactiveness Classifier

There are eleven interactiveness binary classifiers ("Interactiveness Classifier" in Fig. 4) with similar structure in $\mathbf{D}$, i.e., ten for part interactivenesses and one for instance interactiveness. They all take four kinds of features from four aforementioned streams as inputs and are constructed by simple concatenate operation and FC layers. The detailed structure of the interactiveness classifier is illustrated in the upper right part in Fig. 4.

For the **part-level classifier**, the i-th part feature f_{p_i} together with f_h, f_o and f_{sp}, are concatenated and input to FCs and Sigmoids to generate the part interactiveness probability $p^{\mathbf{D}}_{(p_i,o)} = Sigmoid(s^{\mathbf{D}}_{(p_i,o)})$, where $s^{\mathbf{D}}_{(p_i,o)}$ indicates the part interactiveness score of the $i-th$ part. Particularly, fine-grained part interactiveness also has another characteristic, i.e. **sparsity**. For example, in Fig. 6, to judge whether a person is riding a horse, certain parts should be paid more attentions. That is, hands, hip, feet seem more important than head and upper arms. Thus, utilizing part interactiveness as attention is a natural choice to select the important parts, i.e. $f^{'}_{p_i} = p^{\mathbf{D}}_{(p_i,o)} * f_{p_i}$. After re-weighting, the information transported to the next instance-level classifier will be filtered. Thus, model can focus on more important parts and ignore the possible noise imported by the other parts.

0162-8828 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Carleton University. Downloaded on May 26,2021 at 15:21:11 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2021.3054048, IEEE Transactions on Pattern Analysis and Machine Intelligence

For the **instance-level classifier**, we concatenate ten re-weighted part features f'_{p_i} $(1 \leq i \leq 10)$ and input them to FCs to generate the instance interactivenss score $s^{\mathbf{D}}_{(h,o)}$. At last, we obtain the instance interactivenss probability $p^{\mathbf{D}}_{(h,o)} = Sigmoid(s^{\mathbf{D}}_{(h,o)})$. With $p^{\mathbf{D}}_{(h,o)}$ and the binary labels converted from HOI labels, we can construct the binary classification loss $\mathcal{L}^{\mathbf{D}_h}$. Moreover, $s^{\mathbf{D}}_{(h,o)}$ will be used as the final interactiveness score in inference (Sec. 4.4). Notably, we can only obtain the **instance-level** binary labels from instance-level HOI labels, but not the part-level binary labels.

4.3.4 Interactiveness Consistency

With the help of part interactiveness, we can learn deeper interactiveness knowledge via the supervision of *hierarchical consistency*. In detail, a person is interactive if and only if at least one of the body parts is interacting with the object, and is not interactive if and only if none of the parts are interactive. This simple but robust relationship is an important clue for us to mine deeper discriminative information for complex and diverse HOIs. For example, in Fig. 6, Person 1 is riding the horse, which can be inferred by part features because his hands are holding the rope and feet are stepping on the pedals. In contrast, Person 2 and Person 3 are not interacting with the horse because none of their body parts are interacting with the horse.

Theoretically, instance interactiveness is equal to the result of OR operation between all part interactivenesses. In practice, we can use the max pooling to implement the OR operation. This is also in line with the MIL paradigm [14]. Thus, for a human-object pair, the prediction of our method should obey:

$$p^{\mathbf{D}}_{(h,o)} = p^{\mathbf{D}}_{(hp,o)} = max(p^{\mathbf{D}}_{(p_i,o)}), \tag{1}$$

where $p^{\mathbf{D}}_{(p_i,o)}(1 \leq i \leq 10)$ indicates the predicted interactiveness probability for the i-th body part. And $p^{\mathbf{D}}_{(hp,o)}$ is the instance interactiveness probability aggregated from part interactivenesses. We also use $p^{\mathbf{D}}_{(hp,o)}$ to generate another binary classification loss $\mathcal{L}^{\mathbf{D}_{hp}}$. Meanwhile, $p^{\mathbf{D}}_{(h,o)}$ is the instance interactiveness probability from the instance binary classifier (Sec. 4.3.3). $max(\cdot)$ means max pooling operation. In implementation, we use the predicted interactiveness scores to construct the **consistency loss**:

$$\mathcal{L}^{\mathbf{D}_c} = \mathbf{MSE}(s^{\mathbf{D}}_{(h,o)}, max(s^{\mathbf{D}}_{(p_i,o)})), \tag{2}$$

where MSE is the Mean Square Error operation, $s^{\mathbf{D}}_{(h,o)}$ means the instance interactiveness score and $s^{\mathbf{D}}_{(p_i,o)})$ indicates the i-th part interactiveness score. Consistency loss can avoid the conflicts between the knowledge from two different levels and bring stronger supervised guidance.

In conclusion, the loss of interactiveness discriminator $\mathbf{D}$ can be expressed as:

$$\mathcal{L}^{\mathbf{D}} = \mathcal{L}^{\mathbf{D}_h} + \mathcal{L}^{\mathbf{D}_{hp}} + \mathcal{L}^{\mathbf{D}_c}, \tag{3}$$

4.4 Testing with Non-Interaction Suppression

With learned interactiveness knowledge, we can perform Non-Interaction Suppression (NIS) first in testing. Besides, we propose a Low-grade Suppressive Function(LIS) that has the ability to enhance the differentiation between high and low grade object detections. Subsequently, we first discuss LIS and then explain our NIS.

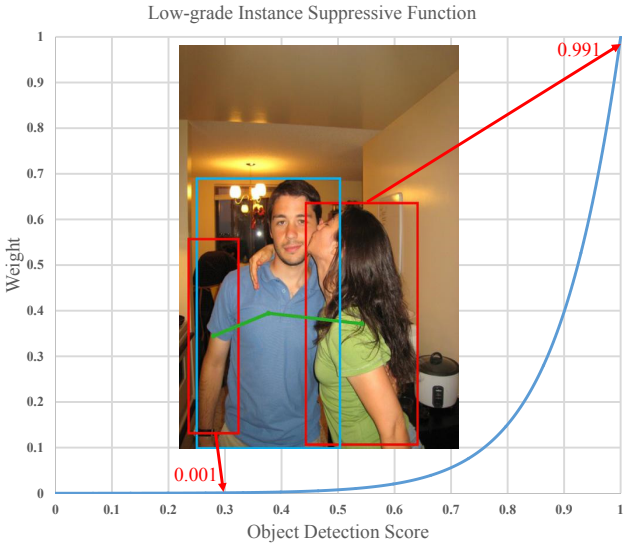

Fig. 7: The illustration of $\mathcal{P}(\cdot)$ within Low-grade Suppressive Function. Its input is object detection score. High-grade detected objects will be emphasized and distinguished with low-grade ones. In addition, $\mathcal{P}(0) = 5.15E-05$ and $\mathcal{P}(1) = 9.99E-01$.

4.4.1 Low-Grade Instance Suppressive Function

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

$$s'^{\mathbf{D}}_{(h,o)} = s^{\mathbf{D}}_{(h,o)} * L(s_h, s_o), \tag{4}$$

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:

$$L(s_h, s_o) = \mathcal{P}(s_h) * \mathcal{P}(s_o), \tag{5}$$

where

$$\mathcal{P}(x) = \frac{T}{1 + e^{(k - wx)}}, \tag{6}$$

$\mathcal{P}(\cdot)$ is a part of the logistic function, the value of $\mathbf{T}$, $\mathbf{k}$ and $\mathbf{w}$ are determined by data-driven manner. Fig. 7 depicts the curve of $\mathcal{P}(\cdot)$ whose domain definition is $(0, 1)$. A bounding box would have low weight till its score is higher than a threshold. Previous works [10], [12] often directly multiply detection scores by the final classification score. But they cannot notably emphasize the differentiation between high quality and inaccurate detection results. In contrast, LIS has the ability to enhance this differentiation as shown in Fig. 7.

4.4.2 Non-Interaction Suppression

After the interactiveness learning, we further utilize $\mathbf{D}$ to suppress the non-interactive pair candidates in **testing**, i.e. Non-Interaction Suppression (NIS). The inference process is based on the tree structure as shown in Fig. 2. Detected instances in the test set will be paired exhaustively, so a dense graph $\mathcal{G}$ of humans and objects is generated. First, we employ $\mathbf{D}$ to compute the interactiveness confidences of all edges. Next, we suppress the edges that meet NIS

0162-8828 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Carleton University. Downloaded on May 26,2021 at 15:21:11 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2021.3054048, IEEE Transactions on Pattern Analysis and Machine Intelligence

conditions, i.e. interactiveness confidence is 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 denoted as:

$$\mathcal{S}^{\mathbf{C}}_{(h,o)} = \mathcal{F}_{\mathbf{C}}[\Gamma^{'}; \mathcal{G}^{'}(v_h, v_o)], \tag{7}$$

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

$$\mathcal{S}_{(h,o)} = \mathcal{S}^{\mathbf{C}}_{(h,o)} * s'^{\mathbf{D}}_{(h,o)}. \tag{8}$$

Here we multiply interactiveness score $s'^{\mathbf{D}}_{(h,o)}$ from $\mathbf{D}$ (Eq. 4) by the output score of $\mathbf{C}$ (Eq. 7).

5 Experiments

In this section, we first introduce the datasets and metrics adopted and then give the implementation details of our framework. Then we introduce two mode settings in the experiment, namely, *Default Joint Learning Mode* and *Transfer Learning Mode*. Next, we report our HOI detection results qualitatively and quantitatively compared with state-of-the-art approaches. Finally, we conduct ablation studies to validate the components in our framework.

5.1 Datasets and Metrics

We mainly adopt three HOI datasets HICO-DET [9], V-COCO [13] and HAKE [15].

HICO-DET [9] includes 47,776 images (38,118 in train set and 9658 in test set), 600 HOI categories on 80 object categories (same with [39]) 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".

HAKE [15] provides 118K+ images, which include 285K human instances, 250K interacted objects, and 724K HOI pairs. The abundance of HOI samples can help our model achieve better performance on interactiveness classification.

HAKE-HOI [15] To better evaluate our method, we re-split HAKE [15] and construct a much larger benchmark: **HAKE-HOI**. It provides 110K+ images (77260 images in the train set, 11298 images in the validation set, and 22156 images in the test set). Compared with HICO-DET [9], HAKE-HOI dataset has larger train set and test set. The interaction categories are similar to the settings of HICO-DET [9], but we exclude the 80 "non-interaction" categories and only define 520 HOI categories. This can help to alleviate the annotation missing in HICO-DET [9].

Metrics. We follow the settings adopted in [9], 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. Additionally, we measure the interactiveness detection in a similar setting. The only difference is that HOI classification is multi-label while the interactiveness classification is binary.

5.2 Implementation Details

We employ a Faster R-CNN [4] with ResNet-50 [37] as $\mathbf{R}$ and keep it frozen. $\mathbf{C}$ consists of three streams similar to [9], [12], extracting features $\mathbf{\Gamma}^{'}$ from instance appearance, spatial location as well as context. Within human and object streams, a residual block [37] with global average pooling and two 1024 sized FCs are used. Relatively, the spatial stream is composed of two convolutional layers with max-pooling, and two 1024 sized FCs. Following [9], [12], we use the late fusion strategy in $\mathbf{C}$.

For interactiveness network $\mathbf{D}$, the human stream, object stream and spatial-pose stream are set the same as in $\mathbf{C}$. For part stream, after pooling, ten part features are first respectively concatenated with the instance-level features from the other three streams, and then passed through two 1024 sized FCs to perform interactiveness discrimination. Whereafter, ten part interactiveness probabilities are generated. Max pooling is then imposed on them to reason out the aggregated instance interactiveness prediction. Meanwhile, ten part features are multiplied by their corresponding interactiveness probabilities for re-weighting. Then, all ten body part features, together with other instance-level features are concatenated again and passed through two 1024 sized FCs to make instance interactiveness prediction. Finally, the consistency loss between two levels is constructed as the objective.

For fair comparison, we adopt the object detection results and COCO [39] pre-trained weights from [12] which are provided by authors. Since NIS and LIS can suppress non-interactive pairs, we set detection confidence thresholds lower than [12], i.e. 0.6 for human and 0.4 for object. The image-centric training strategy [4] is also applied. In other words, pair candidates from one image make up the mini-batch. Cross-validation is used to determine the hyper-parameters. We adopt SGD with cosine decay restart policy and set an initial learning rate as 1e-4, momentum as 0.9. To handle the data bias, we control the ratio of positive and negative pairs for each image as 1:4. We jointly train the framework for 25 epochs. In LIS mentioned in Eq. 6, 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 Default Joint Learning

In *Default Joint Learning Mode*, HOI classifier $\mathbf{C}$ is trained together with $\mathbf{D}$. By adding a supervisor $\mathbf{D}$, 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 overall loss of the proposed method can be expressed as:

$$\mathcal{L} = \mathcal{L}^{\mathbf{C}} + \mathcal{L}^{\mathbf{D}}, \tag{9}$$

where $\mathcal{L}^{\mathbf{C}}$ denotes the HOI classification cross entropy loss, while $\mathcal{L}^{\mathbf{D}}$ is the binary classification cross entropy loss (Eq. 3).

Different from one-stage methods, additional interactiveness discrimination enforces the model to learn interactiveness knowledge, which can bring more powerful constraints. Namely, when a pair is predicted as specific HOIs

0162-8828 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Carleton University. Downloaded on May 26,2021 at 15:21:11 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2021.3054048, IEEE Transactions on Pattern Analysis and Machine Intelligence

such as "cut cake", $\mathbf{D}$ must give the prediction "interactive" simultaneously. Experiment results (Sec. 5.6) prove that interactiveness knowledge learning can effectively refine the training and improve the performance.

An additional benefit of the default joint learning mode is that, if cooperated with the multi-stream HOI detection model $\mathbf{C}$, $\mathbf{D}$ can share the weights of convolutional blocks with the ones in $\mathbf{C}$. To be more specific, blocks $H^{\mathbf{D}}$ and $O^{\mathbf{D}}$ 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 $\mathbf{D}$ and $\mathbf{C}$ in the multi-task training.

The framework in default joint learning mode is called "$\mathbf{RCD}$" in the following, where "$\mathbf{R}$" "$\mathbf{C}$" "$\mathbf{D}$" represents the Representation Network, the Classification Network, and the Interactiveness Network respectively.

5.4 Transfer Learning

As aforementioned, $\mathbf{D}$ only needs binary labels that can be easily converted from HOI labels and are beyond HOI classes. On account of the generalization ability of interactiveness, $\mathbf{D}$ can be trained and transferred across datasets. Therefore, in *Transfer Learning Mode*, $\mathbf{D}$ is used as a transferable knowledge learner to learn interactiveness from multiple datasets, and then be applied to each of them respectively. To be more specific, we train $\mathbf{C}$ and $\mathbf{D}$ **separately** instead of jointly. In transfer learning mode, the benefit of sharing weight is *absent*. However, the independent training makes it possible for $\mathbf{D}$ to learn interactiveness knowledge across datasets, i.e. *flexibility* and *reusability*. Furthermore, the integration of datasets boosts the performance of interactiveness learning. Meanwhile, we can use the cross-trained $\mathbf{D}$ to operate a more powerful NIS in inference.

The mode settings of transfer learning is shown in Tab. 1. The train and test sets of $\mathbf{C}$ are kept the same as default mode, while the train set of $\mathbf{D}$ is adjusted. For transfer learning mode $\mathbf{RCD}_i$, increasingly larger train sets are marked by i from $i = 1$ to $i = 3$, i.e., **1)** V-COCO or HICO-DET, **2)** V-COCO and HICO-DET, **3)** HAKE. That is, $\mathbf{D}$ will be trained with more and more samples. Thus we can analyze the effect of NIS with different scales of interactiveness knowledge.

Meanwhile, different from $\mathbf{D}$, $\mathbf{C}$ 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 transfer abilities of interactiveness knowledge and HOI knowledge in Sec. 5.6.

5.5 Results and Comparisons

We compare our method with five state-of-the-art HOI detection methods [9], [10], [11], [12], [32] on HICO-DET, and four methods [10], [11], [12], [13] on V-COCO. The HOI detection result is evaluated with mAP. For HICO-DET, we follow the settings in [9]: 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) on Scenario 1. For HAKE-HOI, we evaluate mAP following the Default mode of [9] for 520 HOIs. More details can be found in [9], [13].

Test Set	Method	$\mathbf{D}$-Train Set	$\mathbf{C}$-Train Set
HICO-DET	Default $\mathbf{RCD}$	HICO-DET	HICO-DET
	$\mathbf{RCD}_1$	V-COCO	HICO-DET
	$\mathbf{RCD}_2$	HICO-DET, V-COCO	HICO-DET
	$\mathbf{RCD}_3$	HAKE	HICO-DET
V-COCO	Default $\mathbf{RCD}$	V-COCO	V-COCO
	$\mathbf{RCD}_1$	HICO-DET	V-COCO
	$\mathbf{RCD}_2$	HICO-DET, V-COCO	V-COCO
	$\mathbf{RCD}_3$	HAKE	V-COCO
HAKE-HOI	Default $\mathbf{RCD}$	HAKE-HOI	HAKE-HOI

TABLE 1: Mode setting details. Default $\mathbf{RCD}$ is *Default Joint Learning Mode* while $\mathbf{RCD}_i$ is *Transfer Learning Mode*. $\mathbf{RCD}_i$ means that interactiveness network $\mathbf{D}_i$ is trained with increasingly larger datasets marked by i from $i = 1$ to $i = 3$.

	Default			Known Object		
Method	Full	Rare	Non-Rare	Full	Rare	Non-Rare
HO-RCNN [9]	7.81	5.37	8.54	10.41	8.94	10.85
InteractNet [10]	9.94	7.16	10.77	-	-	-
GPNN [11]	13.11	9.34	14.23	-	-	-
iCAN [12]	14.84	10.45	16.15	16.26	11.33	17.73
Peyre et al. [32]	19.40	14.60	20.90	-	-	-
Default $\mathbf{RCD}$	17.84	13.08	18.78	20.58	16.19	21.45
$\mathbf{RCD}_1$	17.49	12.23	18.53	20.28	15.25	21.27
$\mathbf{RCD}_2$	18.43	13.93	19.32	21.10	16.56	22.00
$\mathbf{RCD}_3$	**20.93**	**18.95**	**21.32**	**23.02**	**20.96**	**23.42**
$\mathbf{R}$+iCAN [12]+$\mathbf{D}_3$	17.58	13.75	18.33	19.13	15.06	19.94

TABLE 2: Results comparison on HICO-DET [9].

5.5.1 Results Analysis

Results on HICO-DET and V-COCO are shown in Tab. 2 and Tab. 3 respectively.

Default Joint Learning. The $\mathbf{RC}$ modules in our method adopt similar model design, object detection and backbone with previous method [9], [12]. However, with the joint learning and NIS, our $\mathbf{RCD}$ directly outperforms [9], [12] with 10.03 and 3.00 mAP on HICO-DET. On V-COCO and HAKE-HOI respectively, $\mathbf{RCD}$ also achieves 3.10 mAP and 4.38 mAP improvements compared with iCAN [12]. This greatly verifies the efficacy of hierarchical interactiveness learning.

Transfer Learning. In transfer learning, with the size of $\mathbf{D}$-train set increases, the performance is also improved accordingly. $\mathbf{RCD}_3$ presents the greatest performance improvement and achieves the state-of-the-art performance, i.e., **20.93**, **18.95**, **21.32** mAP on three Default sets on HICO-DET,

Method	AP_{role}
Gupta et al. [13]	31.8
InteractNet [10]	40.0
GPNN [11]	44.0
iCAN w/ late(early) [12]	44.7 (45.3)
Default $\mathbf{RCD}$	48.4
$\mathbf{RCD}_1$	48.5
$\mathbf{RCD}_2$	48.7
$\mathbf{RCD}_3$	**49.1**
$\mathbf{R}$+iCAN w/ late(early) [12]+ $\mathbf{D}_3$	45.8(46.1)

TABLE 3: Results comparison on V-COCO [13].

Method	mAP
iCAN [12]	11.00
Default $\mathbf{RCD}$	**15.38**
$\mathbf{R}$+iCAN [12]+$\mathbf{D}_3$	13.13

TABLE 4: Results comparison on HAKE-HOI.

0162-8828 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Carleton University. Downloaded on May 26,2021 at 15:21:11 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2021.3054048, IEEE Transactions on Pattern Analysis and Machine Intelligence

Test Set	Method	Reduction
HICO-DET	Default **RCD**	-68.73%
HICO-DET	$\mathbf{RCD}_1$	-67.75%
	$\mathbf{RCD}_2$	-74.52%
	$\mathbf{RCD}_3$	-85.93%
V-COCO	Default **RCD**	-50.75%
V-COCO	$\mathbf{RCD}_1$	-59.32%
	$\mathbf{RCD}_2$	-61.55%
	$\mathbf{RCD}_3$	-72.19%
HAKE-HOI	Default **RCD**	-86.20%

TABLE 5: Non-interactive pairs reduction after performing NIS.

Test Set	Method	mAP
HICO-DET	Default **RCD**	14.35
HICO-DET	$\mathbf{RCD}_3$	16.42
HAKE-HOI	Default **RCD**	19.29

TABLE 6: **Interactiveness detection** results comparison on HICO-DET [9] and HAKE-HOI.

and **49.1** mAP on V-COCO. It surpasses the recent state-of-the-art [32] by **1.53**, **4.35**, and **0.42** mAP on three Default sets on HICO-DET. Compared with mode $\mathbf{RCD}_1$, $\mathbf{RCD}_3$ gains an improvement of **3.44** and **2.74** mAP on Default and Known Object Full sets on HICO-DET, and an increase of **6.72** and **5.71** mAP on Rare sets. Noticeably, as the generalization ability of interactiveness is beyond HOI category settings, information scarcity and learning difficulty of rare HOI categories is alleviated. So the performance difference between rare and non-rare HOI categories is accordingly reduced. Our method achieves excellent performance on rare HOIs, the Rare set performance of $\mathbf{RCD}_3$ (18.95 mAP) is even beyond the Non-Rare performances of default **RCD** (18.78 mAP) and previous methods [12] (16.15 mAP). This strongly proves the transferability of interactiveness.

On the other hand, we also apply NIS to iCAN [12] and compare the testing results with the original ones presented in [12]. Results show that NIS also brings an increase of **2.74** and **2.87** mAP on Default and Known Object Full sets on HICO-DET, **1.1** and **0.8** mAP on V-COCO, and **2.13** mAP on HAKE-HOI. This again proves the versatility and effectiveness of the proposed NIS.

Since NIS implicitly evaluates the performance of our interactiveness network, we also *explicitly* measure the interactiveness detection performance via role mean average precision [13] (Tab. 6). Default **RCD** achieves 14.35 and 19.29 mAP respectively on HICO-DET [9] and HAKE-HOI [15]. Benefit from a larger training set, $\mathbf{RCD}_3$ outperforms Default **RCD** by 2.07 mAP on HICO-DET [9].

To further study the relationship between the performance improvement and the training sample scale of **D**, we analyze the non-interactive pairs reduction after employing NIS (Tab. 5). As we cannot annotate all non-interactive pairs in the test set, we compute the ratio according to the assigned binary labels. Specially, HICO-DET has 80 "no-interaction" HOIs correspond to 80 objects. They are "naturally" non-interactive but "positive" for the HICO-DET setting. Therefore, when the NIS condition is satisfied (NIS score is lower than the threshold), we do not discard the pairs belong to these 80 "no-interaction" HOIs and only discard the pairs of other 520 normal HOIs. With interactiveness transferred from more data, $\mathbf{RCD}_3$ achieves the best suppressive effect and discards **85.93%** and **72.19%** non-interactive pairs respectively on two datasets, thus bringing more performance gains. Meanwhile, $\mathbf{RCD}_1$, $\mathbf{RCD}_2$ also perform well and suppresses a certain amount of non-interactive pair candidates. This indicates the great potential of the interactiveness network. Namely, with more training data, **D** is expected to boost the performance further. Moreover, since HICO-DET train set (38K) is much bigger than V-COCO train set (2.5K), the default mode **RCD** achieves larger improvement than $\mathbf{RCD}_1$ on HICO-DET (Tab. 2). And the situation between **RCD** and $\mathbf{RCD}_1$ on V-COCO is opposite. On HAKE-HOI, we also find an effective non-interaction pairs reduction (86.20%).

5.5.2 Visualized Results

Representative predictions are shown in Fig. 8. We find that our model is capable of detecting various 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.

Fig. 8: 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.

Fig. 9 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 a crowd of people with ties. In the bottom-right corner we show an even harder sample. For the HOI "type_on keyboard" between human and the keyboard, **C** wrongly predicts that this HOI is within two close hands. However, **D** 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.

5.5.3 Part Interactiveness Insight

As aforementioned, we use part attention strategy to strengthen the important parts in inference. With $\mathbf{RCD}_3$ (HICO-DET) and **RCD** (HAKE-HOI), we visualize and quantify the part interactiveness patterns for certain HOIs.

0162-8828 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Carleton University. Downloaded on May 26,2021 at 15:21:11 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2021.3054048, IEEE Transactions on Pattern Analysis and Machine Intelligence

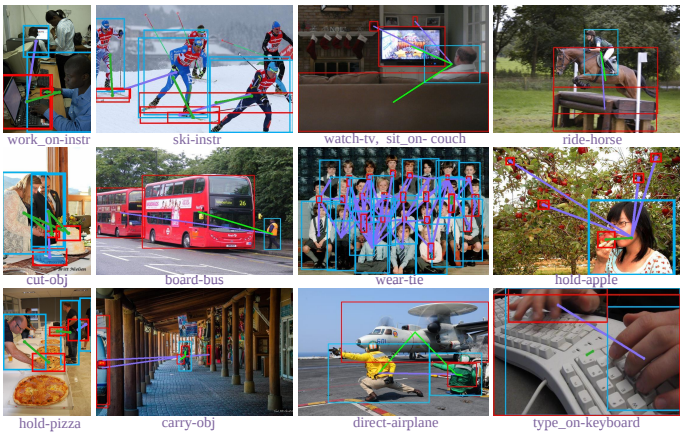

Fig. 9: Visualized effects of NIS. Green lines mean accurate HOIs, while purple lines mean non-interactive pairs which are suppressed. Without NIS, **C** would generate false positive predictions for these non-interactive pairs in one-stage inference, which are shown by the purple texts below the images. 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.

HOI	Feet	Thighs	Hip	Upper arms	Hands	Head
feed cat	0.50	1.00	0.00	0.81	0.52	0.19
pet cat	0.09	1.00	0.62	0.87	0.35	0.00
chase cat	1.00	0.75	0.00	0.58	0.79	0.80
lie-on couch	0.39	1.00	0.61	0.65	0.40	0.00
sit-on couch	1.00	0.89	0.00	0.49	0.40	0.10
board airplane	1.00	0.63	0.00	0.73	0.77	0.66
hold hair-dryer	0.09	1.00	0.51	0.75	0.33	0.00
stand-under stop-sign	0.08	1.00	0.00	0.85	0.43	0.21

TABLE 7: Body parts interactiveness attention pattern of $\mathbf{RCD}_3$ on HICO-DET [9]. For a certain HOI, we first select all the related pairs on HICO-DET [9] and then calculate the average interactiveness scores of each part respectively. Finally, we use a minmax-scaler to rescale the six interactiveness scores to $[0, 1]$ for each HOI, i.e., the highest interactiveness score is scaled to 1.0 while the lowest to 0.0.

Fig. 10 illustrates the heatmaps of interactiveness attentions. The pixels with higher interactiveness probabilities are presented with a brighter red color. We find that part-level interactiveness knowledge localizes the most informative parts effectively. For example, for HOI "swing-baseball_bat" (row2, col5), hands and arms have obvious higher attentions, while hip, thighs and feet are not lightened. With the interactiveness heatmap, we can take further insight into the model. We first compare HOI and non-HOI cases (col1 and col2). For HOI cases, the most possible interactive parts are localized, while body parts all get low interactiveness inference scores for non-HOI cases. Furthermore, we find that the model can learn some functionality of different human body parts rather than simply highlight body parts *physically close* to the object. For the cases "inspect-sports_ball" (row1, col3) and "fly-kite" (row1, col4), body parts "head" and "hands" are respectively highlighted although 10 body parts are equally far away from the object. For the cases "inspect-handbag" (row2, col4), "head" is highlighted instead of the closer "feet" and "legs", which verifies the effectiveness of our interactiveness network. Another similar case is "read-book" (row2, col3).

We also list the quantified results in Tab. 7. For a certain HOI, we first select all the related pairs on HICO-DET and then calculate the average interactiveness scores of each part respectively. For symmetrical parts (feet, thighs, upper arms and hands), the model process the left and right parts separately with different bounding boxes. For simplicity, we average the "left" and "right" scores for the same kind of parts. Finally, for the six parts (feet, thighs, hip, upper arms, hands and head), we use a minmax-scaler to rescale the six interactiveness scores to $[0, 1]$ for each HOI, *i.e.*, the highest interactiveness score is scaled to 1.0 while the lowest to 0.0.

The results in Tab. 7 are insightful. For "cat", different actions have resulted in different part interactiveness patterns. "feed cat" is most related to thighs and upper arms, while "chase cat" is most related to feet, hands and head. This is consistent with common sense. When feeding a cat, a person may let the cat sit on thighs and use arms to give it food. Meanwhile, when chasing a cat, a person may move the feet quickly and try to catch it by hands. Besides, there are also some inaccurate situations. Upper arms and hands are sometimes confused. For "hold hair-dryer", hands should be more interactive than shoulders, but reverse results are generated. For "sit-on couch", the attention of hip is zero, which is unreasonable. Such biases are probably caused by the occlusion. In many training images, the hip is usually invisible and occluded by other objects.

Method	HICO-DET Default Full	V-COCO AP_{role}	HAKE-HOI mAP on Validation Set
iCAN [12]	14.84	45.3	20.45
Default **RCD**	17.84	48.4	26.92
w/o NIS	15.86	46.2	24.86
w/o LIS	16.35	47.4	26.33
w/o NIS & LIS	15.45	45.8	24.57
$\mathbf{RCD}_3$ (Default **RCD** for HAKE-HOI)	20.93	49.1	26.92
w/o part stream	18.52	48.7	25.60
w/o max-pooling	18.85	48.7	25.58
w/o separate binary classifier	20.18	48.9	26.27
$\mathbf{RC}_\mathbf{T}$	10.61	38.5	-

TABLE 8: Results of ablation studies. Following [9], [11], [12], [32], we report the results on HICO-DET and V-COCO test sets. On HAKE-HOI, the results on validation set are shown.

5.6 Ablation Studies

In default **RCD**, we analyze the significance of Low-grade Instance Suppressive and Non-Interaction Suppression in inference (Tab. 8). We also analyze the design of the interactiveness network and the transferability of HOI Knowledge.

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 do not use NIS to discard the non-interactive pairs and directly use the $\mathcal{S}^{\mathbf{C}}_{(h,o)}$ (Eq. 7) as the final prediction. 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 Eq. 4, we observe a degradation in Tab. 8. This suggests that LIS is capable of distinguishing the low-grade detections and improves the performance without using a more costly superior object detector.

NIS & LIS. Without NIS and LIS both, our method only takes effect in the *joint training*. As we can see in Tab. 8, performance degrades greatly but still outperforms iCAN [12], which indicates the enhancement brought by **D** in the hierarchical joint training.

0162-8828 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Carleton University. Downloaded on May 26,2021 at 15:21:11 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2021.3054048, IEEE Transactions on Pattern Analysis and Machine Intelligence

Fig. 10: The heatmaps of interactiveness attention based on $\mathbf{RCD}_3$ on HICO-DET and Default $\mathbf{RCD}$ on HAKE-HOI. The pixels with higher interactiveness probabilities are presented with brighter red color. We can find that part-level interactiveness knowledge localizes the most informative parts effectively. With the interactiveness heatmap, we can take further insight into the model.

Model Design. To verify the components of our model, we train $\mathbf{D}$ without the part stream, i.e., only with three conventional streams: human, object and spatial-pose. We also train our model without the max-pooling stream, i.e., without part-level interactiveness $s^{\mathbf{D}}_{(hp,o)}$ and consistency loss $\mathcal{L}^{\mathbf{D}_c}$. Our model shows obvious degradations in these two situations, which strongly verifies the effectiveness of the hierarchical interactiveness paradigm. Additionally, in our model, we adopt 10 classifiers for 10 parts. Alternatively, we can also train our model with a shared part interactiveness classifier. This modification also degrades the performance, despite slight improvements on computation properties (training/inference speed, model size, parameter). For more details please refer to the supplementary.

Transferability of HOI Knowledge. We also evaluate the transferability of HOIs. $\mathbf{RC}_\mathbf{T}$ means $\mathbf{C}$ is trained on HICO-DET and tested on V-COCO, and vice versa. Compared with iCAN [12], it shows a significant performance decrease of 4.23 and 6.80 mAP on two datasets. On the contrary, $\mathbf{D}$ shows satisfying performance across different datasets. This proves that interactiveness is more suitable and easier to transfer than HOIs knowledge.

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. We propose a hierarchical interactiveness paradigm to adopt both instance and part level interactivenesses. And a consistency learning task is further explored to improve the learning. With interactiveness knowledge, we exploit an interactiveness network to perform Non-interaction Suppression before HOI classification in inference. Extensive experiment results show the efficiency of learned interactiveness knowledge. By combining our method with existing detection models, we achieve state-of-the-art results on the HOI detection task.

Acknowledgment

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 and Shanghai Qi Zhi Institute, SHEITC (2018-RGZN-02046).

References

[1] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma *et al.*, "Visual genome: Connecting language and vision using crowdsourced dense image annotations," *IJCV*, 2017.

[2] C. Lu, R. Krishna, M. Bernstein, and L. Fei-Fei, "Visual relationship detection with language priors," in *ECCV*, 2016.

[3] H.-S. Fang, G. Lu, X. Fang, J. Xie, Y.-W. Tai, and C. Lu, "Weakly and semi supervised human body part parsing via pose-guided knowledge transfer," *CVPR*, 2018.

[4] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *NIPS*, 2015.

[5] C. Lu, H. Su, Y. Li, Y. Lu, L. Yi, C.-K. Tang, and L. J. Guibas, "Beyond holistic object recognition: Enriching image understanding with part states," in *CVPR*, 2018.

[6] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *ICCV*, 2017.

[7] F. Caba Heilbron, V. Escorcia, B. Ghanem, and J. Carlos Niebles, "Activitynet: A large-scale video benchmark for human activity understanding," in *CVPR*, 2015.

[8] B. D. Argall, S. Chernova, M. Veloso, and B. Browning, "A survey of robot learning from demonstration," *Robotics and autonomous systems*, vol. 57, no. 5, 2009.

[9] Y.-W. Chao, Y. Liu, X. Liu, H. Zeng, and J. Deng, "Learning to detect human-object interactions," in *WACV*, 2018.

[10] G. Gkioxari, R. Girshick, P. Dollár, and K. He, "Detecting and recognizing human-object interactions," *arXiv preprint arXiv:1704.07333*, 2017.

[11] S. Qi, W. Wang, B. Jia, J. Shen, and S.-C. Zhu, "Learning human-object interactions by graph parsing neural networks," in *ECCV*, 2018.

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

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

[14] O. Maron and T. Lozano-Pérez, "A framework for multiple-instance learning," in *NIPS*, 1998.

0162-8828 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Carleton University. Downloaded on May 26,2021 at 15:21:11 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2021.3054048, IEEE Transactions on Pattern Analysis and Machine Intelligence

[15] Y.-L. Li, L. Xu, X. Liu, X. Huang, Y. Xu, S. Wang, H.-S. Fang, Z. Ma, M. Chen, and C. Lu, "Pastanet: Toward human activity knowledge engine," in *CVPR*, 2020.
[16] M. A. Sadeghi and A. Farhadi, "Recognition using visual phrases," in *CVPR*, 2012.
[17] M. Yatskar, L. Zettlemoyer, and A. Farhadi, "Situation recognition: Visual semantic role labeling for image understanding," in *CVPR*, 2016.
[18] D. Xu, Y. Zhu, C. B. Choy, and L. Fei-Fei, "Scene graph generation by iterative message passing," in *CVPR*, 2017.
[19] H. Zhang, Z. Kyaw, S.-F. Chang, and T.-S. Chua, "Visual translation embedding network for visual relation detection," in *CVPR*, 2017.
[20] G. Yin, L. Sheng, B. Liu, N. Yu, X. Wang, J. Shao, and C. C. Loy, "Zoom-net: Mining deep feature interactions for visual relationship recognition," *arXiv preprint arXiv:1807.04979*, 2018.
[21] J. Yang, J. Lu, S. Lee, D. Batra, and D. Parikh, "Graph r-cnn for scene graph generation," in *ECCV*, 2018.
[22] Y. Wang, H. Jiang, M. S. Drew, Z.-N. Li, and G. Mori, "Unsupervised discovery of action classes," in *CVPR*, 2006.
[23] W. Yang, Y. Wang, and G. Mori, "Recognizing human actions from still images with latent poses," in *CVPR*, 2010.
[24] N. Ikizler, R. G. Cinbis, S. Pehlivan, and P. Duygulu, "Recognizing actions from still images," in *ICPR*, 2008.
[25] B. Zhuang, Q. Wu, C. Shen, I. Reid, and A. v. d. Hengel, "Care about you: towards large-scale human-centric visual relationship detection," *arXiv preprint arXiv:1705.09892*, 2017.
[26] H.-S. Fang, J. Cao, Y.-W. Tai, and C. Lu, "Pairwise body-part attention for recognizing human-object interactions," in *ECCV*, 2018.
[27] V. Delaitre, I. Laptev, and J. Sivic, "Recognizing human actions in still images: a study of bag-of-features and part-based representations," in *BMVC*, 2010.
[28] Y. W. Chao, Z. Wang, Y. He, J. Wang, and J. Deng, "Hico: A benchmark for recognizing human-object interactions in images," in *ICCV*, 2015.
[29] C.-Y. Chen and K. Grauman, "Predicting the location of "interactees" in novel human-object interactions," in *ACCV*, 2014.
[30] A. Mallya and S. Lazebnik, "Learning models for actions and person-object interactions with transfer to question answering," in *ECCV*, 2016.
[31] L. Shen, S. Yeung, J. Hoffman, G. Mori, and F. F. Li, "Scaling human-object interaction recognition through zero-shot learning," in *WACV*, 2018.
[32] J. Peyre, I. Laptev, C. Schmid, and J. Sivic, "Detecting unseen visual relations using analogies," *arXiv preprint arXiv:1812.05736*, 2018.
[33] G. Gkioxari, R. Girshick, and J. Malik, "Actions and attributes from wholes and parts," in *ICCV*, 2014.
[34] R. Girshick, I. Radosavovic, G. Gkioxari, P. Dollár, and K. He, "Detectron," https://github.com/facebookresearch/detectron, 2018.
[35] T.-Y. Lin, P. Dollár, R. B. Girshick, K. He, B. Hariharan, and S. J. Belongie, "Feature pyramid networks for object detection." in *CVPR*, 2017.
[36] R. Girshick, "Fast r-cnn," in *ICCV*, 2015.
[37] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016.
[38] H.-S. Fang, S. Xie, Y.-W. Tai, and C. Lu, "RMPE: Regional multi-person pose estimation," in *ICCV*, 2017.
[39] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *ECCV*, 2014.

Yong-Lu Li received the B.S. degree in automation from Beijing University of Chemical Technology, Beijing, China, in 2012, and the M.S. degree in Control Engineering from Institution of Automation, Chinese Academy of Science and Harbin University of Science, and Technology, Beijing, China, in 2017. He is currently a Ph.D. candidate with MVIG lab, Shanghai Jiao Tong University. His research interests include human activity understanding, knowledge-based reasoning and robotics.

Xinpeng Liu is currently pursuing the bachelor's degree at Shanghai Jiao Tong University, Shanghai, China. He is a research assistant at Machine Vision and Intelligence Group, Shanghai Jiao Tong University. His research interests include computer vision and deep learning.

Xiaoqian Wu is currently pursuing the bachelor's degree at Shanghai Jiao Tong University, Shanghai, China. She is a research assistant at Machine Vision and Intelligence Group, Shanghai Jiao Tong University. Her research interests include computer vision and deep learning.

Xijie Huang is currently pursuing the bachelor's degree at Shanghai Jiao Tong University, Shanghai, China. He is a research assistant at Machine Vision and Intelligence Group, Shanghai Jiao Tong University. His research interests include computer vision and deep learning.

Liang Xu received the B.S. degree in computer science at Nanjing University, Nanjing, China in 2018. He is currently a second-year master student at Shanghai Jiao Tong University, supervised by Prof. Cewu Lu. His research interests mainly focus on computer vision.

Cewu Lu is a Associate Professor at Shanghai Jiao Tong University (SJTU). Before he joined SJTU, he was a research fellow at Stanford University working under Prof. Fei-Fei Li and Prof. Leonidas J. Guibas. He was a Research Assistant Professor at Hong Kong University of Science and Technology with Prof. Chi Keung Tang. He got the his PhD degree from The Chinese Univeristy of Hong Kong, supervised by Prof. Jiaya Jia. He is one of core technique member in Stanford-Toyota autonomous car project. He serves as an associate editor for Journal gtCVPR and reviewer for Journal TPAMI and IJCV. His research interests fall mainly in Computer Vision, deep learning, deep reinforcement learning and robotics vision.

0162-8828 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Carleton University. Downloaded on May 26,2021 at 15:21:11 UTC from IEEE Xplore. Restrictions apply.