Learning to Detect Human-Object Interactions with Knowledge

Bingjie Xu[1], Yongkang Wong[1], Junnan Li[1], Qi Zhao[2], Mohan S. Kankanhalli[1]
[1]National University of Singapore [2]University of Minnesota
bingjiexu@u.nus.edu, yongkang.wong@nus.edu.sg, lijunnan@u.nus.edu
qzhao@cs.umn.edu, mohan@comp.nus.edu.sg

Abstract

The recent advances in instance-level detection tasks lay a strong foundation for automated visual scenes understanding. However, the ability to fully comprehend a social scene still eludes us. In this work, we focus on detecting human-object interactions (HOIs) in images, an essential step towards deeper scene understanding. HOI detection aims to localize human and objects, as well as to identify the complex interactions between them. Innate in practical problems with large label space, HOI categories exhibit a long-tail distribution, i.e., there exist some rare categories with very few training samples. Given the key observation that HOIs contain intrinsic semantic regularities despite they are visually diverse, we tackle the challenge of long-tail HOI categories by modeling the underlying regularities among verbs and objects in HOIs as well as general relationships. In particular, we construct a knowledge graph based on the ground-truth annotations of training dataset and external source. In contrast to direct knowledge incorporation, we address the necessity of dynamic image-specific knowledge retrieval by multi-modal learning, which leads to an enhanced semantic embedding space for HOI comprehension. The proposed method shows improved performance on V-COCO and HICO-DET benchmarks, especially when predicting the rare HOI categories.

1. Introduction

Recent years have witnessed rapid progress towards visual scene understanding, from object detection [25] to action recognition [23]. However, understanding a scene requires not only detecting individual object instances but also recognizing the visual relationships between object pairs [18, 22, 28]. One particularly important facet of visual relationships is human-centric interaction detection, known as human-object interaction (HOI) detection [5, 10, 13, 15, 33]. Given an input image, it aims to localize all humans and objects, and to identify all the triplets ⟨human, verb, object⟩ (see Figure 1). HOI detection underpins a variety of AI tasks such as visual Q&A [30], robotic task manipulation [3], surveillance event detection [2] and human-centered computing [27]. However, HOI detection is still far from settled due to a large label space of verbs and their interactions with a wide range of object types.

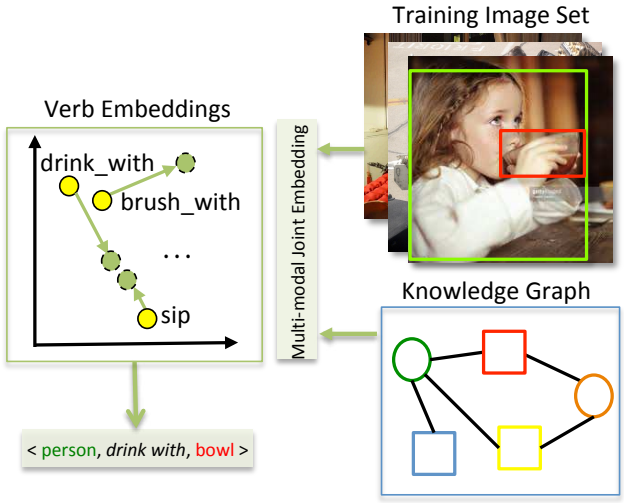

Figure 1: Conceptual illustration of our proposed multi-modal joint embedding learning. In HOI detection task, the label space is often large and intrinsically having long-tail distribution issue where some categories have few samples (*e.g.* ⟨human, drink_with, bowl⟩). The proposed model learns a semantic structure aware embedding space compared to original word embeddings, such that it can leverage semantic similarity to retrieve the verb(s) best describing the detected ⟨human, object⟩ pair. The underlying semantic regularities of verbs and objects are modeled with graph.

Innate in many problems of practical interest, the label space of HOIs that are compositions of humans, verbs and objects exhibits a long-tail distribution, meaning that some categories possess very few training examples. For example, the number of training examples "person riding a bike" is much more than "person riding an elephant". This is a fundamental problem for the standard deep learning approach, which relies on amount of data for each category to obtain an effective discriminative pattern. Though HOIs are

visually diverse, the compositional elements (human, verb, object) contain intrinsic semantic regularities [14]. Specifically, verbs and objects in the HOIs share certain characteristics across various types of scenes. For example, the similar shape of "bike" and "elephant", as well as the similar spatial configuration of "on top of" have implications for the verb "ride" and its semantic-close verb "sit_on". Motivated by this observation, we propose to learn to detect HOIs by modeling the underlying regularities among the verbs and object categories in visual relationships using a graph based approach.

Existing works [19, 47] have attempted to tackle the long-tail distribution issue in HOI recognition with multi-modal sources. However, these works do not explicitly consider the joint impact from the referent source and visual information. In contrast, our work emphasizes on the joint update of the verb embeddings from vision and linguistic knowledge by introducing a multi-modal embedding module to dynamically learn the semantic dependencies of the concepts. The idea of joint update is illustrated in Figure 1. In the original word embedding space, "drink_with" is mapped to be close to "brush_with" possibly because of the shared "with" from word vector embeddings [32]. With the joint update from knowledge graph and training image set, the model is more likely to comprehend the meaning of "drink" as its neighbors include "sip".

In this work, we aim to answer two essential questions in leveraging knowledge to enhance the HOI detection task with long-tail label distribution: (1) how to model the semantic regularities of knowledge for HOIs? and (2) how to dynamically retrieve the image-specific associated knowledge? Our proposed model solves them by: first, mining structure of verbs and object categories from internal training annotations and external annotations of general visual relationships dataset [28]; second, introducing a multi-modal verb embedding space with joint update from visual representations and linguistic knowledge. The contributions are summarized as follows:

- In order to address the long-tail distribution issue in HOI detection, we construct a knowledge graph to model the dependencies of the verbs and object categories in HOIs and other visual relationships.
- We offer a new perspective into HOI detection with multi-modal embeddings such that the model can learn the associated verb expression refering to its semantic structure for each visual query.
- We achieve improved performance on two benchmarks, especially for rare HOI categories, and conduct extensive ablation study to identify the relative contributions of the individual components. Our code will be publicly available[1].

[1]https://bitbucket.org/freezingmolly/hoi_graph

2. Related Work

HOI Understanding. Different from general visual relationships, which focuses on two arbitrary objects in the images, HOIs are human-centric with fine-grained verb-object labels. HOI understanding starts from the concept of "affordance" [11]. The fine-grained understanding has been scaled up with the advances from deep learning and several large-scale HOI datasets [1, 5, 6, 15, 47]. Works have been done for learning to detect HOIs with constraints from interacting object locations [13, 15], pairwise spatial configuration [5] to scene context of instances [10, 33]. Another stream of work addresses the long-tail HOI problem with compositional learning [38] and extra image supervision [47]. Our work is in the similar vein to address the long-tail problem in HOI detection, but leverages semantic regularities from linguistic knowledge. The proposed multi-modal verb embedding explicitly considers the reciprocity between referent knowledge and visual information.

Learning with Long-Tail Labels. The *instrinsic* long-tail property of label space poses challenges in realistic tasks [4]. In the context of triplet detection with long-tail labels, considering a triplet as an unique class is a hindrance for scalability. Therefore, compositional learning [19, 38, 44] has been employed to learn each compositional label in the triplet that is less rare individually. Modeling with the internal data has attempted to link the classes in head to tail [41, 45]. Works [9, 19, 28, 35, 39, 42, 43, 46, 47] have also attempted to exploit external sources of the same or different modalities complementing the examined data to boost learning from very few examples. Our work follows compositional learning and leverages semantic regularities from linguistic knowledge, but in the context of HOI detection that verbs and interacting objects are linked with semantic structures.

Graph Neural Networks. Some approaches have been proposed that apply deep neural networks to graph structured data. One group of approaches applies feed-forward neural networks to every node of the graph recurrently, for example Graph Neural Network (GNN) [37] and an improved version Gated Graph Neural Network (GGNN) [24]. The second group is to generalize convolution layers to the graphs. In the direction of spectral approach that requires spectral representations of graph structure, Graph Convolutional Network (GCN) [20] is proposed for semi-supervised learning to process language. In contrast, non-spectral approaches operate convolutions directly on the graphs [8, 16]. Graph neural networks have been exploited to model scene instance dependencies [7, 22, 33] and knowledge structures [19, 31, 40]. Our model is similar to the direction on modeling knowledge structures, but exploits GCN to jointly match the verb semantic embeddings rather than verb-object categories to visual representation, to enhance comprehension of the verbs compositionally.

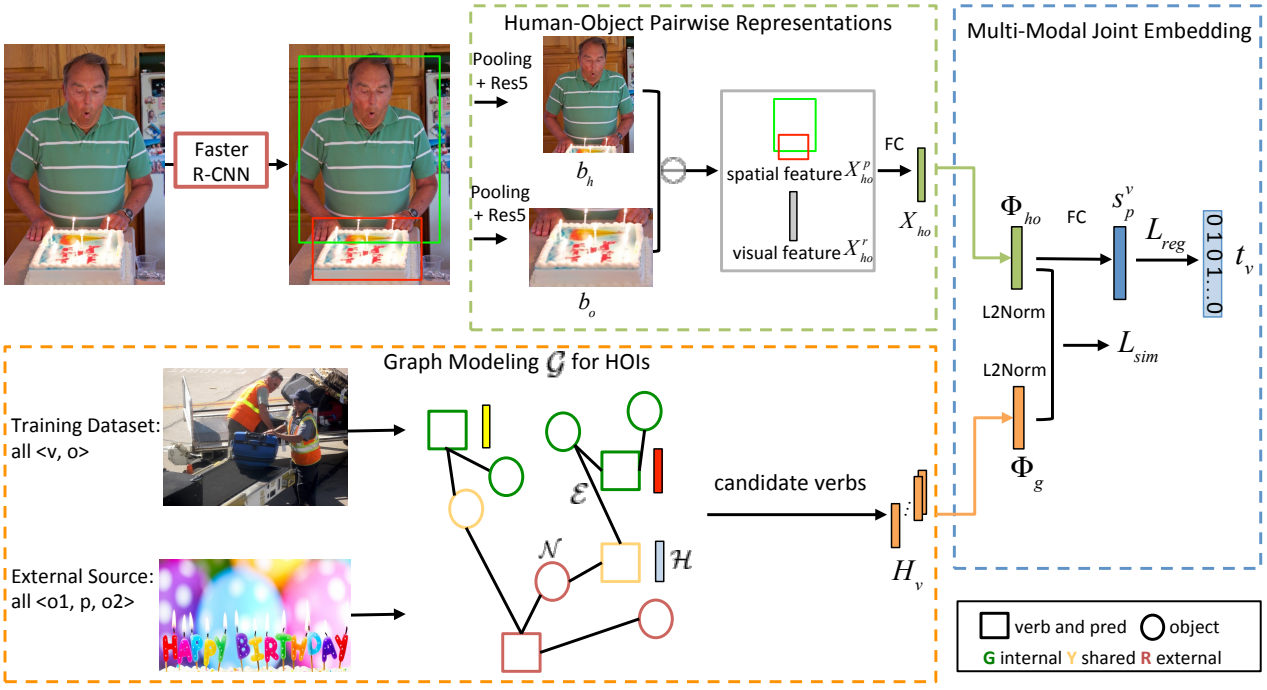

Figure 2: Illustration of the proposed verb embedding learning, which consists of human-object visual representation module (Section 3.3), knowledge graph modeling module (Section 3.2), and joint embedding module (Section 3.4). FC and $\ominus$ indicate fully-connected layer and element-wise subtraction, respectively. Respective predictions based on human and object feature vectors are not included for clarity.

3. Method

The task of HOI detection is to detect humans and objects in an image, as well as identify verbs of each $\langle$human, object$\rangle$ pair. During the training phase, training data consists of ground-truth HOI annotations, labeled as $\langle$human, verb, object$\rangle$ triplets, where the human and objects are localized as bounding boxes. Given a learned model and a given probe image, the inference process predicts all possible verbs based on the detected humans and objects.

3.1. Problem Formulation

Formally, given a set of human and object regions $b_h, b_o \in \mathcal{B}$ proposed for an image I, a set of triplet score $S^v_{h,o}$ is assigned to each $\langle b_h, b_o\rangle$ pair representing the probability of verbs and the pair detection. Unlike learning of verb-object categories as a whole with complexity of $\mathcal{O}(|\mathcal{V}| \cdot |\mathcal{C}|)$, where $|\mathcal{V}|$ and $|\mathcal{C}|$ are respectively the numbers of verbs and object categories, we decompose $S^v_{h,o}$ to enable combinations of all verbs and objects for a complexity of $\mathcal{O}(|\mathcal{V}| + |\mathcal{C}|)$:

$$S^v_{h,o} = s_h \cdot s_o \cdot (s^v_h \cdot s^v_o \cdot s^v_{h,o}) \tag{1}$$

where s_h (s_o) is the human (object) detection class score of b_h (b_o), s^v_h (s^v_o) is the verb prediction score from human (object) stream, and $s^v_{h,o}$ is the verb prediction score from the human-object pairwise representation. Please see Figure 3 for the detailed inference procedure.

We propose a novel method to calculate $s^v_{h,o}$ that incorporates knowledge graph to address the long-tail problem in class distributions. Different from existing HOI detection approaches [5, 10, 13, 33], we formulate HOI detection as a verb retrieval task with a given visual query pair $\langle b_h, b_o\rangle$. Formally, we maximize the likelihood of the conditional distribution of the referent verb(s) $v^* \in \mathcal{V}$ as:

$$v^* = \arg\max_{v \in \mathcal{V}} p(v | X_h, X_o, \mathcal{G}) \tag{2}$$

where $\mathcal{G}$ is the relational knowledge graph which incorporates linguistic information available from training dataset and external source (Section 3.2). X_h and X_o are respectively the visual representation for b_h and b_o (Section 3.3).

To tackle the formulated verb retrieval task, we project the visual and linguistic information to a joint verb embedding space such that the embeddings for matched $\langle b_h, b_o\rangle$ and v^* pairs are closer while the unmatched pairs are far away. Figure 2 illustrates our proposed verb embedding learning. The inclusion of knowledge graph allows the joint verb embeddings to exploit the reciprocal nature of referent expression and visual information. Thus, considering the joint verb embedding space as a hidden variable ϕ (detailed in Section 3.4), the likelihood in Eq. 2 can be written as:

$$\sum_{\phi} p(v, \phi | X_h, X_o, \mathcal{G}) = \sum_{\phi} \underbrace{p(v | \phi, X_h, X_o, \mathcal{G})}_{\text{Inference}} \underbrace{p(\phi | X_h, X_o, \mathcal{G})}_{\text{Joint Embedding}} \tag{3}$$

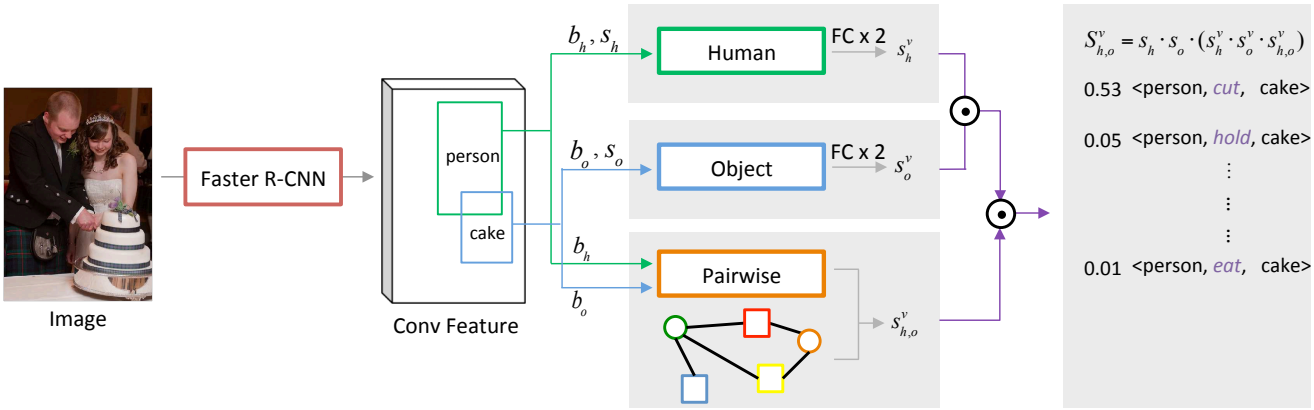

Figure 3: The inference procedure of the proposed model. Given an input image, the proposed model detects HOI triplets and outputs the triplet scores. Element-wise multiplication ($\odot$) is applied to human stream's verb prediction s_h^v, object stream's verb prediction s_o^v and the pairwise verb prediction score $s_{h,o}^v$.

3.2. Graph Modeling for HOIs

In order to boost learning of long-tail classes, we exploit a graph-based approach to model the semantic dependencies from linguistic knowledge complementing the training visual representations.

3.2.1 Preliminary: Graph Convolutional Network

Given the word embeddings of verbs and object categories, the goal of graph modeling is to update the node representations based on ground-truth relations. Formally, we define the knowledge graph as $\mathcal{G} = (\mathcal{N}, \mathcal{E}, \mathcal{H})$, where $\mathcal{N}$ are the nodes, $\mathcal{E}$ are undirected edges linking pairs of nodes, and $\mathcal{H}$ represents the feature vectors of nodes.

To model the semantic dependencies of verbs and object categories, we construct a knowledge graph based on Graph Convolution Network (GCN) [20], which is originally proposed for semi-supervised entity classification. The core idea of GCN is to transform the node features based on the neighboring nodes defined by the adjacency matrix. Mathematically, given a graph adjacency matrix A and node features $H \in \mathcal{H}$, the convolutional operations for the k-th layer in GCN is represented as:

$$H^{k+1} = D^{-\frac{1}{2}}\tilde{A}D^{-\frac{1}{2}}H^kW^k, \text{ where } \begin{cases} \tilde{A} = A + I \\ D_{ii} = \sum_j \tilde{A}_{ij} \end{cases} \quad (4)$$

where A is normalized by the diagonal node degree matrix D with self-connections. $H^k \in \mathbb{R}^{|\mathcal{N}| \times d_k}$ is the input feature vector and $H^{k+1} \in \mathbb{R}^{|\mathcal{N}| \times d_{k+1}}$ is the output feature vector. d_k and d_{k+1} is dimension of the the input and output feature vector. $W^k \in \mathbb{R}^{d_k \times d_{k+1}}$ is the weight matrix specific to the k-th layer, operating on each node feature H^k. The convolutional layers are usually stacked multiple times. A non-linear operation, such as the $\mathrm{ReLU}(\cdot) = \max(0, \cdot)$ can be applied to the output of each convolutional layer.

3.2.2 Graph Convolutional Network for HOIs

When learning the semantically-meaningful node features H, we employ the links of nodes to learn W in each layer. Specifically, the graph structure is used to capture the semantic dependencies amongst verb and object categories in HOIs, and general visual relationships. Formally, nodes $\mathcal{N}$ model all possible verbs and object categories in the annotations, represented by word embeddings of verbs $H_v \in \mathcal{H}$ and objects $H_o \in \mathcal{H}$ from GloVe model [32]. Each undirected edge from $\mathcal{E}$ connects a valid pair of verb and object category according to the $\langle$verb, object$\rangle$ annotations from training dataset, and $\langle$object1, predicate, object2$\rangle$ triplets from general visual relationships dataset [28]. The tail verb classes are thus impacted from its neighbors on the same object node. The intuition of incorporating general visual relationships supplementary to HOIs, such as *preposition* and *spatial* configurations, is that they intrinsically connect to verbs. For example, "a person on the bike" has implications of "person riding the bike".

The adjacency matrix A is initialized with binary values defining the connections (or disconnections) of nodes. The knowledge graph here is task-oriented with manageable size and computational cost. Specifically, the knowledge graph for V-COCO dataset consists of 226 vertices whereas HICO-DET dataset has 313 vertices (both with visual relationships). For both datasets, the number of undirected edges starting from each vertice is less than 193. The node-level update can be shared in parallel at each layer.

3.3. Visual Representations

Given a detected person b_h and a detected object b_o, the learned pairwise representations should preserve their semantic interactions. For example, the interaction (*e.g. sit*) can be characterized by the visual appearance of $\langle b_h, b_o \rangle$ (*e.g. human pose, object shape and size*), and the relative

location configuration. Therefore, for either b_h or b_o, the feature vector X_h or X_o is the concatenation of visual feature X^r of the region from feature extraction backbone, and spatial configuration $X^p = [\frac{x-x'}{w'}, \frac{y-y'}{h'}, \log\frac{w}{w'}, \log\frac{h}{h'}]$.

Here, x and y are the coordinates of the region with size $w \times h$, whereas x', y', w' and h' are of the other region in the pair. Inspired by visual translation embedding [44], we perform subtraction operation on X^r and X^p of $b_h(b_o)$, followed by two respective fully-connected layers, to extract the pairwise representations X^p_{ho} and X^r_{ho}. The final pairwise representation X_{ho} is obtained by concatenating the visual and spatial features followed by one 512 sized fully-connected layer as:

$$X_{ho} = FC(X^p_{ho} \circ X^r_{ho}) \tag{5}$$

3.4. Multi-Modal Joint Embedding Learning

The proposed joint embedding learning aims to distill information from semantic dependencies to jointly learn an embedding for HOI detection. Specifically, the goal is to learn the transformations of visual feature $f_{ho}(X_{ho}) \rightarrow \phi_{ho}$ and GCN feature $f_g(H_v) \rightarrow \phi_g$, such that the learned pairwise embedding of $\langle b_h, b_o \rangle$ can preserve the semantic structure of verbs. This approach guides the learning of verb embeddings by exploiting the semantic regularities associated with visual modality and knowledge.

The objective of the joint embedding learning is to maximize the similarity between positive $\langle \phi_{ho}, \phi_g \rangle$ pairs, and minimize it between all non-matching pairs to a specified margin, as well as preserve the discriminative ability. To this end, we use a combination of similarity loss $\mathcal{L}_{sim}$ [36], cross entropy regularization loss $\mathcal{L}_{reg}$ and cross entropy loss $\mathcal{L}_{cls}$ from individual streams.

Similarity Loss. $\mathcal{L}_{sim}$ for each $\langle \phi_{ho}, \phi_g \rangle$ pair is defined as:

$$\mathcal{L}_{sim}(\phi_{ho}, \phi_g, t_{sim}) = \begin{cases} 1 - \cos(\phi_{ho}, \phi_g), & t_{sim} = 1 \\ \max(\cos(\phi_{ho}, \phi_g) - \alpha, 0), & t_{sim} = 0 \end{cases} \tag{6}$$

where α is the margin. If $\langle b_h, b_o \rangle$ and v is associated in ground-truth, the label t_{sim} is assigned to be 1, otherwise 0.

Cross Entropy Regularization Loss. $\mathcal{L}_{reg}$ is applied for verb classification from the pairwise verb embedding, defined as cross entropy loss on the predicted verb scores $s^v_p \in \mathbb{R}^{|\mathcal{V}|}$. The probabilities are obtained from a shared fully-connected layer applied on ϕ_{ho}, followed by a sigmoid activation to simultaneously predict verb scores. We assign multi-class verb labels t_v based on the ground-truth.

Cross Entropy Loss. $\mathcal{L}_{cls}$ is individually applied to human stream and object stream. The respective X_h and X_o are passed through two fully-connected layers and sigmoid classifiers to obtain verb prediction scores s^v_h and s^v_o. Then, we compute $\mathcal{L}_{cls}$ between s^v_h (s^v_o) and the ground-truth verb labels t_v.

Therefore, the final loss function can be obtained as:

$$\begin{aligned} \mathcal{L} = \lambda_1 \mathcal{L}_{sim}(\phi_{ho}, \phi_g, t_{sim}) + \lambda_2 \mathcal{L}_{reg}(s^v_p, t_v) \\ + \lambda_3 \mathcal{L}_{cls}(s^v_h, s^v_o, t_v) \end{aligned} \tag{7}$$

where λ_1, λ_2 and λ_3 are weights to control the contribution of each loss term. $t_v \in \mathbb{R}^{|\mathcal{V}|}$ denote the labels for the visual modality. Maximizing the joint embedding term in Eq. 3 is equivalent to minimizing Eq. 7.

During the inference stage (see Figure 3), for each $\langle b_h, b_o \rangle$ pair, the pairwise prediction score $s^v_{h,o}$ is obtained from the regularized verb score $s^v_p \cdot \mathsf{softmax}(\cos(\phi_{ho}, \phi_g))$. Thus the triplet score $S^v_{h,o}$ is obtained according to Eq. 1.

4. Experiments

In this section, we first describe the evaluated benchmark datasets (*i.e.* V-COCO [15] and HICO-DET [5]), the evaluation metric, and the implementation details. We also compare our proposed model with the state-of-the-art models, and conduct ablation studies to examine the proposed knowledge modeling and multi-modal embeddings.

4.1. Datasets and Metrics

Dataset. In this work, we evaluate our model on two benchmarks for HOI detection. First, the **V-COCO dataset** [15] is a subset of MS-COCO [26], with 5,400 images in the train-val (training plus validation) set and 4,946 images in the test set. It is annotated with 26 unique verb classes, and has bounding boxes for humans and interacting objects. In particular, three verb classes (*i.e.* cut, hit, eat) are annotated with two types of targets (*i.e.* instrument and direct object). Second, the **HICO-DET dataset** [5] contains 38,118 images in the training set and 9,658 test images, annotated with 600 types of interactions: 80 MS-COCO object categories and 117 unique verbs. The bounding boxes of humans and corresponding objects are also annotated.

Evaluation Metrics. We follow the standard evaluation metric and report role mean average precision (role mAP). mAP is computed based on both recall and precision, which is appropriate for the detection task. The goal is to correctly detect all of the $\langle \mathsf{human}, \mathsf{verb}, \mathsf{object} \rangle$ triplets for an image. A triplet is considered as a true positive if (1) the predicted triplet label is the same as the ground-truth, and (2) both the predicted human and object bounding boxes have intersection-over-union (IoU) greater than 0.5 w.r.t the ground-truth annotations.

4.2. Implementation Details

For fair comparison, we use Faster R-CNN [34] with ResNet-50 [17] as the feature extraction backbone. The pre-trained weight for MS-COCO [26] is from [10]. Human and object bounding boxes are detected with ResNet-50-FPN [25] backbone as [10]. Human and object bounding

Table 1: Comparisons with the state-of-the-art approaches on HICO-DET dataset [5]. Mean average Precision (mAP) (%) for the default setting (object unknown) is reported where higher values indicates better performance. The best scores are marked in **bold**.

Method	Feature Backbone	Full ↑	Rare ↑	Non-Rare ↑
Random	-	1.35e-3	5.72e-4	1.62e-3
Fast-RCNN [12]	CaffeNet	2.85	1.55	3.23
HO-RCNN [5]	CaffeNet	7.81	5.37	8.54
Shen *et al*. [38]	VGG-19	6.46	4.24	7.12
VSRL [13, 15]	ResNet-50-FPN	9.09	7.02	9.71
InteractNet [13]	ResNet-50-FPN	9.94	7.16	10.77
GPNN [33]	ResNet-152	13.11	9.34	14.23
iCAN [10]	ResNet-50	12.80	8.53	14.07
Ours	ResNet-50	**14.70**	**13.26**	**15.13**

boxes with detection confidence scores above 0.8 and 0.4 respectively are kept. Through grid-search on the validation set, the hyper-parameters are set as $\lambda_1 = 0.8$, $\lambda_2 = 1$, and $\lambda_3 = 1$. The margin α for cosine loss is set as 0.1. A mini-batch consists of one positive sample, the jittering positives and negative samples. The negative samples are obtained by pairing all the detected humans and objects that are not annotated in the ground-truth labels, such that the model can learn the pairwise patterns for the negative $\langle \mathsf{human}, \mathsf{object} \rangle$ pairs. We use Stochastic Gradient Descent (SGD) to train the model for 450k iterations with a learning rate of 0.001, a weight decay of 0.0005, and a momentum of 0.9.

Each person can perform multiple verbs on the same object simultaneously, therefore binary sigmoid classifiers are employed for multilabel verb classification. We then minimize the binary cross entropy losses between the ground-truth labels and the predicted scores. Note that in HICO-DET, simultaneous verbs need to be manually combined for each pair of $\langle \mathsf{human}, \mathsf{object} \rangle$ based on IoU of bounding boxes due to separate verb annotations.

To obtain the word embeddings for GCN node inputs, we use the GloVe text model [32] trained on the Wikipedia dataset, which leads to vectors of $\mathbb{R}^{1\times 300}$. For the classes whose names contain multiple words, we empirically average all matched words embeddings. The graph consists of two layers, both with dimension of 512. LeakyReLU with negative slope of 0.2 [40] is used as the activation after each layer of the graph.

4.3. Results

Baselines. We compare our method with the following baselines: (1) Fast-RCNN [12]: predictions are obtained by linearly combining the human and object detection scores. (2) HO-RCNN [5]: a multi-stream model combines the scores from appearance of human and object, as well as spatial configuration of the pair. (3) VSRL [15]: uses spatial constraints for the interacting objects. We report the reimplemented result from [13]. (4) Shen *et al*. [38]: predictions are from separate verb and object training. (5) InteractNet [13]: multi-loss of object detection, human-object pairwise prediction, and additional human-centric branch that learns a human action-specific density function. (6) BAR-CNN [21]: a modified Faster R-CNN detection pipeline augmented with a box attention mechanism. (7) GPNN [33]: a dynamic scene graph based approach with node outputs as verb and object predictions. (8) iCAN [10][2]: a multi-stream model of human, object appearance and pairwise spatial configuration with additional attention based context.

Table 2: Comparisons with the state-of-the-art approaches on V-COCO dataset [15]. mAP (role) (%) is evaluated as in the standard evaluation metric. Higher values are better. The best scores are marked in **bold**.

Method	Feature Backbone	Sce. 1 ↑
VSRL [15, 13]	ResNet-50-FPN	31.8
InteractNet [13]	ResNet-50-FPN	40.0
BAR-CNN [21]	Inception-ResNet	41.1
GPNN [33]	ResNet-152	44.0
iCAN [10]	ResNet-50	45.3
Ours	ResNet-50	**45.9**

Experiment Results. We present the overall quantitative results on V-COCO (Table 2) and HICO-DET (Table 1). We observe that our proposed model achieves competitive results over the state-of-the-art approaches [13, 33, 10]. For V-COCO, we follow the original evaluation protocol [15]. Compared to the best performing model iCAN, we achieve an absolute gain of $+0.6$. For HICO-DET, we can also observe consistent improvements on all, rare, and non-rare HOI splits against existing best performing methods [33, 10]. We achieve absolute gains of $+1.59$, $+3.92$,

[2]mAP(%) on three category sets for HICO-DET in arxiv version: 14.84, 10.45, 16.15.

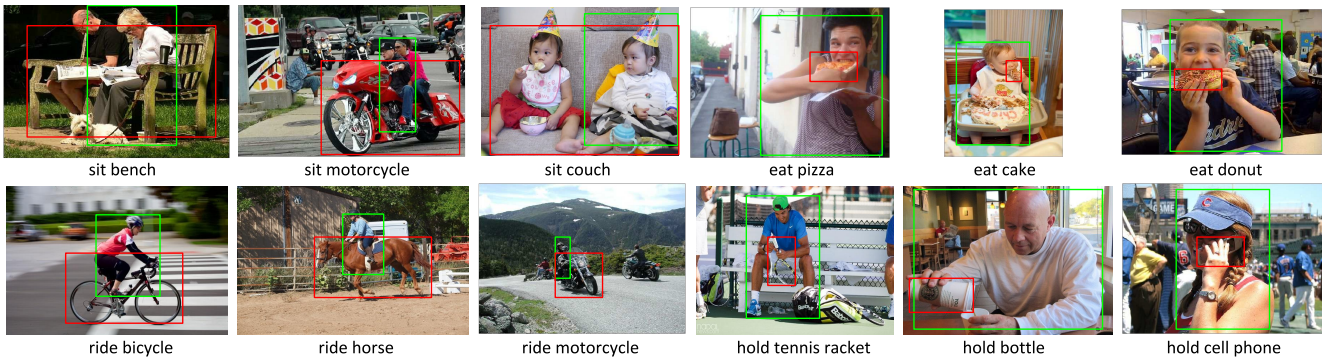

Figure 4: Prediction samples on V-COCO *test* set (first row) and HICO-DET *test* set (second row). Our model detects same type of verbs with various object categories in different scenes, as well as different types of HOIs with the same kind of object. The prediction with the highest HOI triplet score is displayed.

and $+0.9$ over GPNN, respectively.

Figure 4 shows sample HOI detection results on both datasets. We highlight the top-1 prediction result in each image. It shows that our model is capable of predicting verbs interacting with various types of objects, as well as different verbs on the objects of the same category.

4.4. Ablation Study

We analyze the contributions of various components of our model. Table 3 shows the results on both benchmarks.

Extra Knowledge. We first examine the influence of exploiting knowledge in complementing the visual information. We directly combine verb predictions with X_h, X_o, X_{ho} from human, object and pairwise stream, respectively (full model *w/o knowledge*). For recognizing the challenging rare categories, the results are in favor of the whole version of model than *w/o knowledge*. In this case, the model has to use the extra knowledge for less common verbs and objects combinations. This result supports our core argument - extra knowledge about the semantic dependencies can be used to improve HOI predictions especially for long-tail HOI categories. External knowledge of visual relationships is also tested to validate its contribution on modeling dependencies of general predicates-object structures.

Graph Modeling. We here examine the influence of modeling semantic dependencies based on a graph. We directly feed the word2vec verb features, which are originally used for GCN node inputs, into two fully-connected layers to obtain verb embeddings (denoted as *w/o graph*). We can observe that the performance of *w/o graph* is worse than the whole model. This indicates that modeling semantic dependencies of verbs-objects in relationships and leveraging message passing capabilities of GCNs together is essential. Embeddings for pairs of nodes with edges between them impact each other more than the distant ones.

Table 3: Ablation study of our model on V-COCO and HICO-DET test set. Mean average precision (mAP) (%) are reported.

Method	V-COCO	HICO-DET		
	Sce.1 ↑	Full ↑	Rare ↑	None-Rare ↑
Ours	**45.9**	**14.70**	**13.26**	**15.13**
w/o knowledge	42.4	12.55	10.21	13.25
w/o joint embed.	44.0	13.12	11.59	13.58
w/o graph	44.1	13.16	11.63	13.62
w/o external set	45.1	13.91	12.52	14.33

Joint Embedding. We also examine the influence from the proposed multi-modal knowledge retrieval. We concatenate each node output vector of verbs in either V-COCO or HICO-DET with X_{ho}. An averaged vector is obtained by averaging on all concatenated multimodal vectors, and passed through three fully connected layers to get the verb prediction in the pairwise stream. Final predictions are combinations of predictions from human, object and pairwise streams (denoted as *w/o joint embedding*). The decrease in performance supports the effect from joint update of the verb embeddings, which preserve the classification ability and the structural semantics.

Qualitative Ablations. We provide qualitative results in Figure 5. Specifically, we compare the prediction results between the whole model and *w/o knowledge*. It helps to understand the benefits of extra knowledge. Given the same image, our full model (the first row) is more confident to detect the less seen HOIs such as "flip skateboard" based on semantic similarity to "jump skateboard". However, only given the visual information *w/o knowledge*, the model is limited in predicting the concurrent HOIs confidently such as "hold/swing/wield baseball bat".

4.5. Analysis of the Learned Embeddings

To gain further insight into the learned verb embeddings, we explore whether the embedding space has certain en-

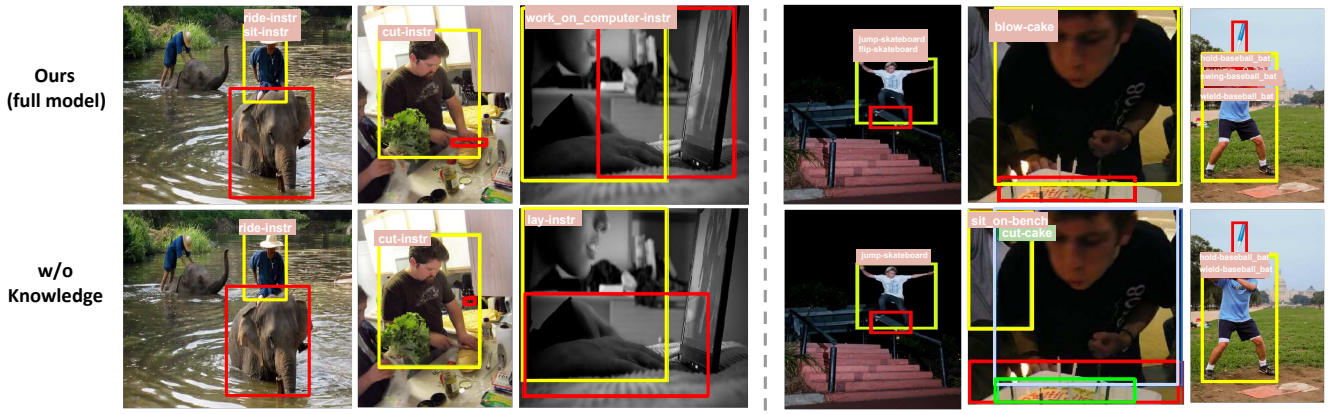

Figure 5: Example of detection results from *Our full model* (first row) and *model w/o knowledge* variant (second row). The first three columns from left show detections on V-COCO *test* set and the remaining columns are from HICO-DET *test* set. Predictions with HOI triplet score > 0.2 are displayed, and "no_interaction" class is not displayed for clarity. Text is annotated with the same color as the corresponding object bound box.

hanced clustering properties in Figure 6 with t-SNE visualization [29]. We show the t-SNE plots of both the word embeddings (input to GCN) and the updated verb embeddings of 117 verbs in HICO-DET dataset. By inspecting the semantic affinities between the embeddings in Figure 6 (a) and (b), we can observe that the original GloVe embeddings without the proposed joint embedding space yields a less accurate projection of data. For examples, GloVe projects "drink_with" to be close to "brush_with" possibly due to the shared "with". However, after the joint update with visual samples and knowledge graph, the model is more likely to understand the meaning of "drink_with" as its neighbors include "sip" and "eat". The semantic dependencies of verbs on an object is also learned, *e.g.* "jump" and "flip" the skateboard. This observation explains the contribution from multi-modal verb embedding learning.

5. Conclusion

In this paper, we aimed to tackle the long-tail distribution issue in the label space for human-object interaction (HOI) categories, which is currently not effectively resolved in HOI detection task. Towards this challenge, we dynamically retrieved the associated linguistic knowledge by introducing a multi-modal embedding space and relational graph. This joint embedding space explicitly considers the cooperative impact between pairwise visual information and associated subgraph of knowledge. We then implemented with image-specific knowledge retrieval. We evaluated our model on two HOI detection benchmarks, and showed promising results. Moving forward, we can address challenges such as understanding human behavior with implications from HOI detection. Specifically, human interactions may imply their intent, and possibly provide information about the past or future thus to help describing various

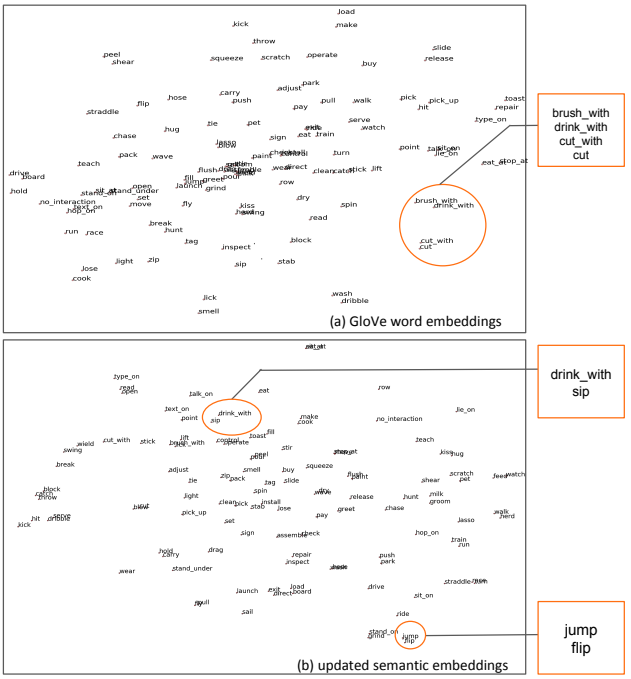

Figure 6: Visual illustration of 117 verb embeddings in HICO-DET dataset via t-SNE visualization [29]. Top is the GloVe word embeddings [32] and the bottom is the semantic embeddings learned with proposed method.

dynamic behavior series. Learning knowledge from noisy web information to alleviate ambiguity in HOI comprehension can also be explored.

Acknowledgment

This research is supported by the National Research Foundation, Prime Minister's Office, Singapore under its Strategic Capability Research Centres Funding Initiative.

References

[1] PIC: Person in context. http://picdataset.com/challenge/index/.

[2] Amit Adam, Ehud Rivlin, Ilan Shimshoni, and David Reinitz. Robust real-time unusual event detection using multiple fixed-location monitors. *IEEE TPAMI*, 30(3):555–560, 2008.

[3] Brenna Argall, Sonia Chernova, Manuela M. Veloso, and Brett Browning. A survey of robot learning from demonstration. *Robotics and Autonomous Systems*, 57(5):469–483, 2009.

[4] Samy Bengio. Sharing representations for long tail computer vision problems. In *ICMI*, page 1, 2015.

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

[6] Yu-Wei Chao, Zhan Wang, Yugeng He, Jiaxuan Wang, and Jia Deng. HICO: A benchmark for recognizing human-object interactions in images. In *ICCV*, pages 1017–1025, 2015.

[7] Ching-Yao Chuang, Jiaman Li, Antonio Torralba, and Sanja Fidler. Learning to act properly: Predicting and explaining affordances from images. In *CVPR*, pages 975–983, 2018.

[8] David K. Duvenaud, Dougal Maclaurin, Jorge Aguilera-Iparraguirre, Rafael Gómez-Bombarelli, Timothy Hirzel, Alán Aspuru-Guzik, and Ryan P. Adams. Convolutional networks on graphs for learning molecular fingerprints. In *NIPS*, pages 2224–2232, 2015.

[9] Andrea Frome, Gregory S. Corrado, Jonathon Shlens, Samy Bengio, Jeffrey Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. Devise: A deep visual-semantic embedding model. In *NIPS*, pages 2121–2129, 2013.

[10] Chen Gao, Yuliang Zou, and Jia-Bin Huang. iCAN: Instance-centric attention network for human-object interaction detection. In *BMVC*, page 41, 2018.

[11] James J Gibson. *The ecological approach to visual perception: classic edition*. Psychology Press, 2014.

[12] Ross Girshick. Fast R-CNN. In *ICCV*, pages 1440–1448, 2015.

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

[14] E Bruce Goldstein and James Brockmole. *Sensation and perception*. Cengage Learning, 2016.

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

[16] William L. Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. In *NIPS*, pages 1025–1035, 2017.

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[18] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Judy Hoffman, Fei-Fei Li, and C. Lawrence Zitnick Ross Girshick. Inferring and executing programs for visual reasoning. In *ICCV*, pages 2989–2998, 2017.

[19] Keizo Kato, Yin Li, and Abhinav Gupta. Compositional learning for human object interaction. In *ECCV*, pages 247–264, 2018.

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

[21] Alexander Kolesnikov, Christoph H. Lampert, and Vittorio Ferrari. Detecting visual relationships using box attention. *arXiv preprint arXiv:1807.02136*, 2018.

[22] Junnan Li, Yongkang Wong, Qi Zhao, and Mohan S. Kankanhalli. Dual-glance model for deciphering social relationships. In *ICCV*, pages 2669–2678, 2017.

[23] Junnan Li, Yongkang Wong, Qi Zhao, and Mohan S. Kankanhalli. Unsupervised learning of view-invariant action representations. In *NeurIPS*, pages 1262–1272, 2018.

[24] Yujia Li, Daniel Tarlow, Marc Brockschmidt, and Richard S. Zemel. Gated graph sequence neural networks. In *ICLR*, 2015.

[25] Tsung-Yi Lin, Piotr Dollár, Ross B. Girshick, Kaiming He, Bharath Hariharan, and Serge J. Belongie. Feature pyramid networks for object detection. In *CVPR*, pages 936–944, 2017.

[26] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *ECCV*, pages 740–755, 2014.

[27] Zhenguang Liu, Zepeng Wang, Luming Zhang, Rajiv Ratn Shah, Yingjie Xia, Yi Yang, and Xuelong Li. Fastshrinkage: Perceptually-aware retargeting toward mobile platforms. In *ACM Multimedia*, pages 501–509, 2017.

[28] Cewu Lu, Ranjay Krishna, Michael S. Bernstein, and Fei-Fei Li. Visual relationship detection with language priors. In *ECCV*, pages 852–869, 2016.

[29] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of machine learning research*, pages 2579–2605, 2008.

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

[31] Kenneth Marino, Ruslan Salakhutdinov, and Abhinav Gupta. The more you know: Using knowledge graphs for image classification. In *CVPR*, pages 20–28, 2017.

[32] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *EMNLP*, pages 1532–1543, 2014.

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

[34] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *NIPS*, pages 91–99, 2015.

[35] Fereshteh Sadeghi, Santosh Kumar Divvala, and Ali Farhadi. VisKE: Visual knowledge extraction and question answering by visual verification of relation phrases. In *CVPR*, pages 1456–1464, 2015.

[36] Amaia Salvador, Nicholas Hynes, Yusuf Aytar, Javier Marín, Ferda Ofli, Ingmar Weber, and Antonio Torralba. Learning

cross-modal embeddings for cooking recipes and food images. In *CVPR*, pages 3068–3076, 2017.

[37] Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. The graph neural network model. *IEEE Transactions on Neural Networks*, 20(1):61–80, 2009.

[38] Liyue Shen, Serena Yeung, Judy Hoffman, Greg Mori, and Fei-Fei Li. Scaling human-object interaction recognition through zero-shot learning. In *WACV*, pages 1568–1576, 2018.

[39] Qian Wang and Ke Chen. Zero-shot visual recognition via bidirectional latent embedding. *International Journal of Computer Vision*, 124(3):356–383, 2017.

[40] Xiaolong Wang, Yufei Ye, and Abhinav Gupta. Zero-shot recognition via semantic embeddings and knowledge graphs. In *CVPR*, pages 6857–6866, 2018.

[41] Yu-Xiong Wang, Deva Ramanan, and Martial Hebert. Learning to model the tail. In *NIPS*, pages 7032–7042, 2017.

[42] Xun Xu, Timothy M. Hospedales, and Shaogang Gong. Semantic embedding space for zero-shot action recognition. In *ICIP*, pages 63–67, 2015.

[43] Ruichi Yu, Ang Li, Vlad I. Morariu, and Larry S. Davis. Visual relationship detection with internal and external linguistic knowledge distillation. In *ICCV*, pages 1068–1076, 2017.

[44] Hanwang Zhang, Zawlin Kyaw, Shih-Fu Chang, and Tat-Seng Chua. Visual translation embedding network for visual relation detection. In *CVPR*, pages 3107–3115, 2017.

[45] Xiao Zhang, Zhiyuan Fang, Yandong Wen, Zhifeng Li, and Yu Qiao. Range loss for deep face recognition with long-tailed training data. In *ICCV*, pages 5419–5428, 2017.

[46] Bohan Zhuang, Lingqiao Liu, Chunhua Shen, and Ian D. Reid. Towards context-aware interaction recognition for visual relationship detection. In *ICCV*, pages 589–598, 2017.

[47] Bohan Zhuang, Qi Wu, Chunhua Shen, Ian D. Reid, and Anton van den Hengel. HCVRD: A benchmark for large-scale human-centered visual relationship detection. In *AAAI*, pages 7631–7638, 2018.