# ConsNet: Learning Consistency Graph for Zero-Shot Human-Object Interaction Detection

Ye Liu
Wuhan University, China
ye-liu@whu.edu.cn

Junsong Yuan
State University of New York at Buffalo, USA
jsyuan@buffalo.edu

Chang Wen Chen
Peng Cheng Laboratory, China
The Chinese University of Hong Kong, Shenzhen, China
State University of New York at Buffalo, USA
chencw@buffalo.edu

## ABSTRACT

We consider the problem of Human-Object Interaction (HOI) Detection, which aims to locate and recognize HOI instances in the form of $\langle human, action, object \rangle$ in images. Most existing works treat HOIs as individual interaction categories, thus can not handle the problem of long-tail distribution and polysemy of action labels. We argue that multi-level consistencies among objects, actions and interactions are strong cues for generating semantic representations of rare or previously unseen HOIs. Leveraging the compositional and relational peculiarities of HOI labels, we propose *ConsNet*, a knowledge-aware framework that explicitly encodes the relations among objects, actions and interactions into an undirected graph called *consistency graph*, and exploits Graph Attention Networks (GATs) to propagate knowledge among HOI categories as well as their constituents. Our model takes visual features of candidate human-object pairs and word embeddings of HOI labels as inputs, maps them into visual-semantic joint embedding space and obtains detection results by measuring their similarities. We extensively evaluate our model on the challenging V-COCO and HICO-DET datasets, and results validate that our approach outperforms state-of-the-arts under both fully-supervised and zero-shot settings.

## CCS CONCEPTS

• **Computing methodologies → Activity recognition and understanding**; *Scene understanding*.

## KEYWORDS

Human-Object Interaction Detection, Graph Neural Networks, Zero-Shot Learning

**ACM Reference Format:**
Ye Liu, Junsong Yuan, and Chang Wen Chen. 2020. ConsNet: Learning Consistency Graph for Zero-Shot Human-Object Interaction Detection. In *Proceedings of the 28th ACM International Conference on Multimedia (MM '20), October 12–16, 2020, Seattle, WA, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3394171.3413600


ACM ISBN 978-1-4503-7988-5/20/10…$15.00
https://doi.org/10.1145/3394171.3413600

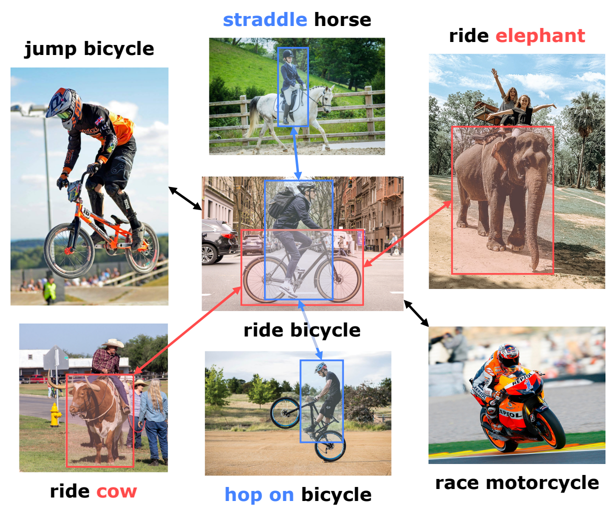


**Figure 1: Illustration of knowledge-aware human-object interaction detection. Red, blue and black lines represent functionally similar objects, behaviorally similar actions and holistically similar interactions. We argue that successful detection of an HOI should benefit from the knowledge obtained from similar objects, actions and interactions.**

## 1 INTRODUCTION

Beyond detecting individual human or object instances in images, it is crucial for machines to also recognize how they interact with each other, which can be essential cues to understand the human-centric visual world. The task of Human-Object Interaction (HOI) Detection aims to locate and recognize HOI instances in images. For example, detecting $\langle human, feed, cat \rangle$ refers to locating "human" and "cat", as well as predicting the action "feed" for this human-object pair. Instead of inferring ambiguous spatial relations among objects, e.g. "cat is on the bed", HOI detection plays a pivotal role to understand *what is happening* in the scene. Studying HOIs can benefit many down-stream visual understanding tasks including image captioning [24], image retrieval [44] and visual question answering [12].

Most existing works on HOI detection [9, 11, 14, 25, 36, 39, 41] treat HOIs as individual interaction categories and focus on mining visual representations of human-object pairs to improve classification performances. Despite previous successes, these conventional

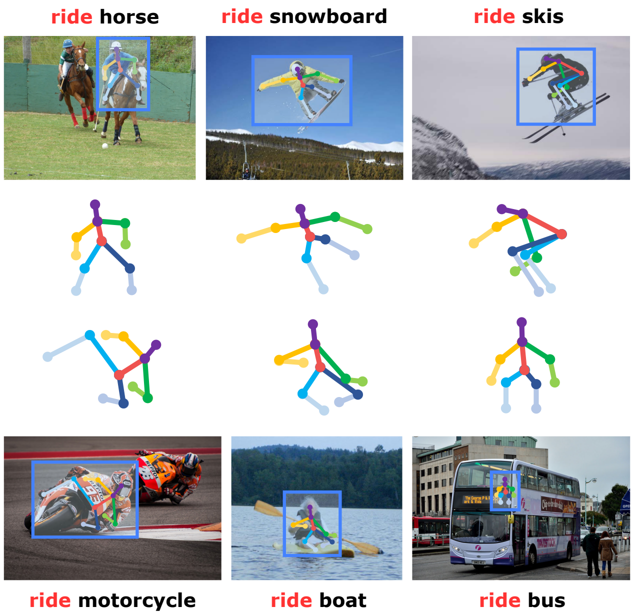


**Figure 2: Polysemy of action labels. All the HOIs above share the same action label "ride", but the actual implications of these actions are inconsistent, as can be seen from the inferred human poses.**

approaches still face two challenges. First, compared with other action-based recognition tasks, what makes HOI detection challenging is that labels of HOIs are fine-grained and are related to the specific object category. The quadratic number of combinations of actions and objects brings prohibitive annotation cost. Hence, non-compositional methods [4, 9, 25, 34, 39, 41] are largely restricted by the coverage and long-tail distribution of exhaustive HOI annotations. Second, the compositional peculiarity of HOI labels also leads to the polysemy of action labels. As an example shown in Figure 2, collocated with different objects, the actual implications of action "ride" are sometimes inconsistent. Such phenomenon brings ambiguities and extra challenges to compositional methods [1, 11, 14, 36].

In this work, we address the above two challenges by proposing a knowledge-aware approach (as shown in Figure 1) for HOI detection. For the first challenge, we claim that the key to dealing with the imbalance and scarcity of HOI training samples is to distill knowledge obtained from non-rare categories, and transfer it to rare or unseen ones. Considering that humans have the ability to perceive unseen interactions, e.g. $\langle human, ride, elephant\rangle$, because they can make use of their common sense to *imagine* what it would be like based on similar HOIs such as $\langle human, ride, bicycle\rangle$ and $\langle human, feed, elephant\rangle$, as well as similar actions or objects such as "sit on" or "horse". To jointly capture the compositional peculiarities and multi-level similarities among HOIs, we define three types of consistencies at different granularities. At unigram level, we introduce **functional consistency**, which depicts the functional similarities among objects, and **behavioral consistency**, which represents the similarities of human behavior when performing different actions. At trigram level, we present **interactional consistency**, which denotes the holistic similarities among HOIs. We further construct an undirected graph, namely **consistency graph**, to explicitly encode these relations. Each node in the consistency graph represents an HOI label or its entities. The three types of consistencies are encoded as edges among the nodes. That is, two object (also action or interaction) nodes are linked if they have whichever the consistencies above. We then use word embeddings of HOI labels as input features of nodes, and exploit recently introduced Graph Attention Networks (GATs) [38] to perform message passing on the consistency graph, enabling the model to learn semantic representations of HOIs in a transductive manner.

When it comes to the second challenge, we argue that an appropriate perception of HOI should benefit from both unigram and trigram representations. Take detecting the HOI $\langle human, ride, bicycle\rangle$ for instance. At unigram level, we ought to make sure that the subject is a human, the object is a bicycle and the subject is performing the action "ride". At trigram level, we should also deem that the human-object pair is performing the right interaction holistically. In our model, HOI detection scores are estimated based on the similarities between visual and semantic embeddings of human, object, action and interaction. Such a decomposition strategy helps capture implications of HOIs at multiple granularities, thus can better handle the polysemy of action labels. Moreover, our model has the ability to transfer knowledge from familiar HOIs, and detect HOIs with unseen actions, objects, or action-object combinations. Note that detecting HOIs with unseen actions may not be performed by previous methods.

The main contributions of our work are as follows:

- We propose a knowledge-aware approach to model relations among HOIs at both unigram and trigram level, and exploit Graph Attention Networks to predict semantic representations of HOIs based on their word embeddings.
- We introduce a data-driven method to estimate consistencies and construct the consistency graph using visual-semantic representations of HOI labels, which can jointly capture visual and semantic features of HOIs.
- Our approach outperforms state-of-the-arts under both fully-supervised and zero-shot settings on the challenging V-COCO and HICO-DET datasets. Further experiments also show that our model has the ability to detect HOIs with unseen actions, which may not be performed by previous methods.

## 2 RELATED WORKS

*Human-Object Interaction Detection.* Human-Object Interaction Detection plays a crucial role in human-centric scene understanding since the problem was first introduced by Gupta and Malik [13]. Most previous works can be divided into compositional methods [1, 11, 14, 36] and non-compositional methods [4, 9, 25, 34, 39, 41]. Compositional methods learn separate detectors for objects and actions. The final detection results are generated by fusing the confidences of three HOI entities. However, these approaches can not handle the polysemy of action labels. Non-compositional methods avoid this problem by predicting HOI labels at the trigram level directly, but they are largely restricted by the coverage and long-tail distribution of HOI annotations. A recently introduced hybrid

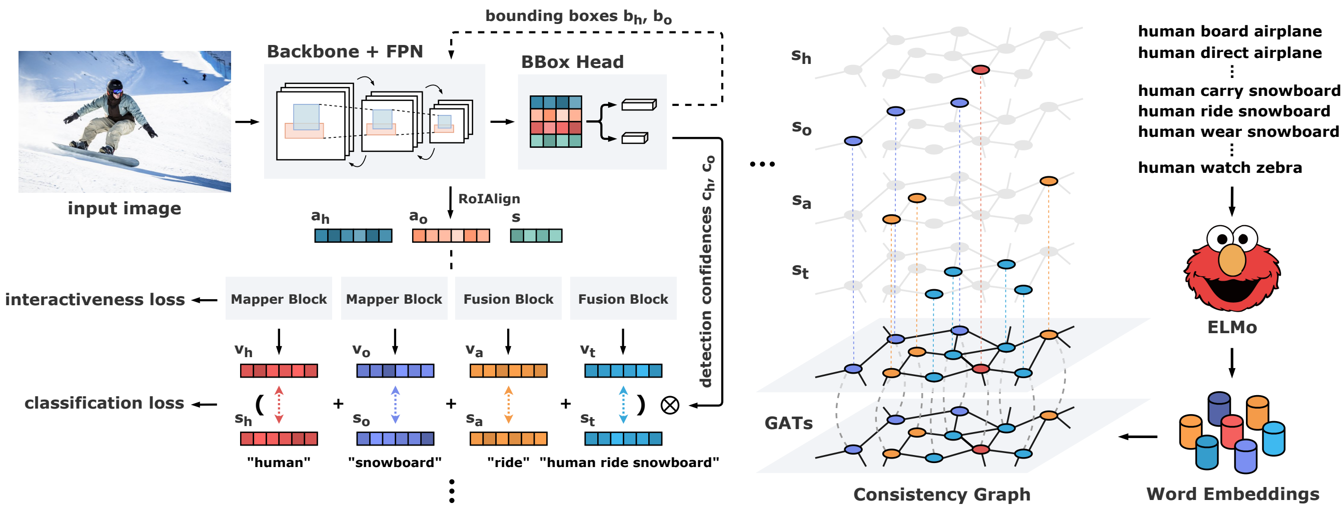


**Figure 3: Overall architecture of our framework. The input image is fed into a pre-trained object detector to obtain bounding boxes $b_h, b_o$ with detection confidences $c_h, c_o$ of humans and objects. The bounding boxes are then used to crop visual features $a_h, a_o$ from FPN and compute spatial configuration $s$. Subsequently, visual embedding network maps these features into multi-level visual embeddings $v_h, v_o, v_a, v_t$. On the other side, semantic embedding network encodes HOI labels into vectors using a pre-trained language model. The word embeddings serve as input features of nodes in the consistency graph. By performing GATs, these features are propagated among neighboring nodes and be transformed into semantic embeddings $s_h, s_o, s_a, s_t$. The HOI detection results are then generated by measuring the similarities among visual embeddings and semantic embeddings.**

model [33] have shown that using both unigram and trigram representations of HOIs may solve the above contradiction. Nonetheless, all these methods ignore the implicit relations among HOI categories, thus we extend the hybrid model by aggregating common sense knowledge for generating semantic embeddings.

*Graph Neural Networks.* The past few years have witnessed the rapid development of representation learning on graphs [46]. The majority of these methods are under the Message Passing Neural Networks (MPNN) framework [10] which decomposes the pipeline into message functions, vertex update functions and readout function. Kipf *et al.* [20] proposed GCNs that extend convolution operation [22] from euclidean data to non-euclidean data. Wu *et al.* [43] simplified GCNs by removing the non-linearities and merging the weights. Hamilton *et al.* [15] proposed GraphSAGE to realize inductive learning on graphs. In this work, we exploit the recently introduced GATs that incorporate multi-head attention mechanism to model the relations of neighboring nodes. The learned attention coefficients in GATs can serve as the weights of consistencies.

*Zero-Shot Learning.* Most recent zero-shot learning approaches can be divided into two protocols [42]. One is to learn semantic representations of categories that can be mapped to visual classifiers [2, 3, 45]. The other is to make use of knowledge graphs to distill the knowledge [7, 8, 29]. In this work, with the help of GNNs and language models, we learn the explicit and implicit knowledge of HOIs from consistency graph and word embeddings, respectively.

# 3 APPROACH

In this section, we introduce our approach on knowledge-aware HOI detection. As illustrated in Figure 3, the entire framework can be divided into two sub-modules, namely visual embedding network and semantic embedding network. These sub-modules can map visual representations of human-object pairs and word embeddings of HOI labels into visual-semantic joint embedding space. HOI detection results are then generated by measuring similarities among visual and semantic embeddings.

## 3.1 Overview

Given an image $x$ and a set of HOI categories of interest $\mathcal{H} = \{1, ..., C\}$, the task of human-object interaction detection is to detect all the human-object pairs in $x$, where the humans and objects are participating one or multiple pre-defined interactions. The outputs of HOI detection would be a set of tuples $\mathcal{T} = \{\langle b_h, b_o, y_{h,o} \rangle\}$, where $b_h, b_o \in \mathbb{R}^4$ denotes bounding boxes of the human and the object, and $y_{h,o}$ represents a vector where $y^i_{h,o} \in \{0, 1\}$ indicates whether the HOI class $i$ is assigned to this human-object pair. Note that a person may have several interactions with multiple objects simultaneously, thus different HOIs may share the same human, action or object.

We adopt a three-stage HOI detection pipeline by generating a set of human-object pairs as candidates, filtering out non-interactive candidates and classifying the remaining ones into multiple interaction categories. In the first stage, an object detector is used to collect bounding boxes of humans $B_h$ and objects $B_o$, as well as their corresponding detection confidences $C_h, C_o$. We only keep top $N_k$ detection results with confidences $c_k$ higher than a threshold $\theta_k$, where $k \in \{h, o\}$ indicates human or object. The candidates are then obtained by pairing up all the remaining humans and objects extensively.

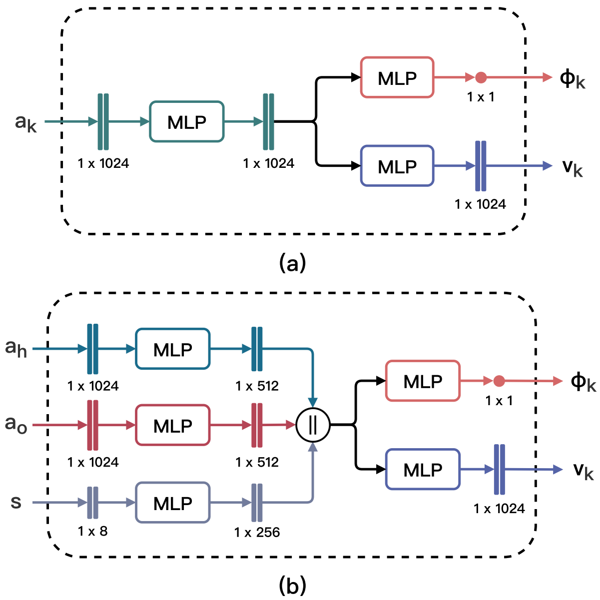


**Figure 4: Detailed architecture of visual embedding network. a) Mapper block takes visual features of human or object $a_k, k \in \{h, o\}$ as inputs, and predicts visual embeddings $v_k, k \in \{h, o\}$ as well as interactiveness $\varphi_k, k \in \{h, o\}$. b) Fusion block takes human features $a_h$, object features $a_o$ and their spatial configuration $s$ as inputs, and estimates visual embeddings of action or interaction $v_k, k \in \{a, t\}$, together with interactiveness $\varphi_k$.**

Recent works have shown that in most cases, the majority of humans and objects in an image are not interacting with each other. Such a severe imbalance between positive and negative candidates makes HOI classification challenging. To address this problem, Li *et al.* [25] proposed the strategy of non-interactive suppression (NIS) to filter out and suppress potential non-interactive candidates. In the second stage, we predict the class-irrelevant interactiveness $\varphi_{h,o}$ for each candidate by

$$\varphi_{h,o} = \sigma(\sum_{k \in \{h,o,a,t\}} \varphi_k) \tag{1}$$

where $\sigma(\cdot)$ denotes the Sigmoid function and $\varphi_k, k \in \{h, o, a, t\}$ indicates the interactiveness score at human, object, action or interaction level. Candidates with interactiveness $\varphi_{h,o}$ lower than a threshold $\theta_{h,o}$ would be discarded. The remaining ones are then fed into HOI classifier for further interaction classification.

In the third stage, we classify the candidates into HOI categories in a knowledge-aware manner. For each candidate, the confidence of assigning HOI class $i$ to it can be given by

$$P(y_i = 1 | x, b_h, b_o, c_h, c_o) = r^i_{h,o} \cdot \varphi_{h,o} \cdot c_h \cdot c_o \tag{2}$$

where $r^i_{h,o}$ is the HOI classification score given by the HOI classifier. Interactiveness $\varphi_{h,o}$, human detection confidence $c_h \in C_h$ and object detection confidence $c_o \in C_o$ serve as suppression terms on potential non-interactive or non-existent candidates. The HOI classification score $r^i_{h,o}$ can be given by

$$r^i_{h,o} = \sigma(\sum_{k \in \{h,o,a,t\}} \frac{v_k \cdot s^i_k}{\|v_k\|_2 \cdot \|s^i_k\|_2} \cdot \gamma) \tag{3}$$

where $v_k$ denotes visual embeddings of the candidate, including human $v_h$, object $v_o$, action $v_a$ and interaction $v_t$. $s^i_k$ represents semantic embeddings of these entities for HOI class $i$. We treat $s_k$ as templates of HOIs and measure the distance among visual and semantic embeddings by computing cosine similarities. Note that we also add a scale factor $\gamma$ to control the range of outputs.

The visual embeddings $v_k$, interactiveness $\varphi_{h,o}$ and semantic embeddings $s_k$ are generated by visual embedding network and semantic embedding network. Details of the embedding networks are explained in the following sections.

## 3.2 Visual Embedding Network

Visual embedding network takes image $x$ as well as bounding boxes of human and object $b_h, b_o$ as inputs, and generates visual embeddings of human $v_h$, object $v_o$, action $v_a$ and interaction $v_t$. These visual embeddings are constructed based on visual features of human $a_h$, object $a_o$ and their spatial configuration $s$. We adopt ResNet-50-FPN [17, 27], which can be shared with the object detector, as the feature extractor. We obtain the visual features of human and object by cropping the appropriate level of feature map from FPN using RoIAlign [16] according to their bounding boxes. Spatial configuration of a candidate is computed by

$$s = \big\|_{k \in \{h,o\}} (\frac{x^k_1 - d_x}{\psi} \| \frac{x^k_2 - d_x}{\psi} \| \frac{y^k_1 - d_y}{\psi} \| \frac{y^k_2 - d_y}{\psi}) \tag{4}$$

where $\|$ denotes concatenation operation, $x^k_1, x^k_2, y^k_1, y^k_2, k \in \{h, o\}$ are coordinates of the human or object bounding box, $(d_x, d_y)$ and $\psi$ represent the origin and area of the union box respectively. The computed spatial configuration $s$ would be a $1 \times 8$ vector. We hypothesize that visual embeddings of human and object can be predicted by their own visual features $a_k, k \in \{h, o\}$, while visual embeddings of action and interaction are jointly affected by visual features of human and object $a_k, k \in \{h, o\}$ as well as their spatial configuration $s$.

$$P(\varphi_m, v_m \mid x, b_h, b_o) = P(\varphi_m, v_m \mid a_m), m \in \{h, o\} \tag{5}$$

$$P(\varphi_n, v_n \mid x, b_h, b_o) = P(\varphi_n, v_n \mid a_h, a_o, s), n \in \{a, t\} \tag{6}$$

Based on the hypotheses above, we introduce two types of embedding blocks, i.e. mapper block and fusion block, to predict interactiveness $\varphi_{h,o}$ and generate visual embeddings $v_k, k \in \{h, o, a, t\}$ for candidates. Details of the embedding blocks are described in section 3.2.1 and 3.2.2.

### 3.2.1 *Mapper Block.*

As shown in Figure 4 (a), mapper block only takes visual features of the human or object as inputs. These visual features are first transformed into hidden states by a multi-layer perceptron (MLP). After that, two MLPs are used to map the dimensions of hidden states to $1 \times 1$ and $1 \times 1024$ respectively. The two outputs are interactiveness $\varphi_k$ and visual embeddings $v_k$.

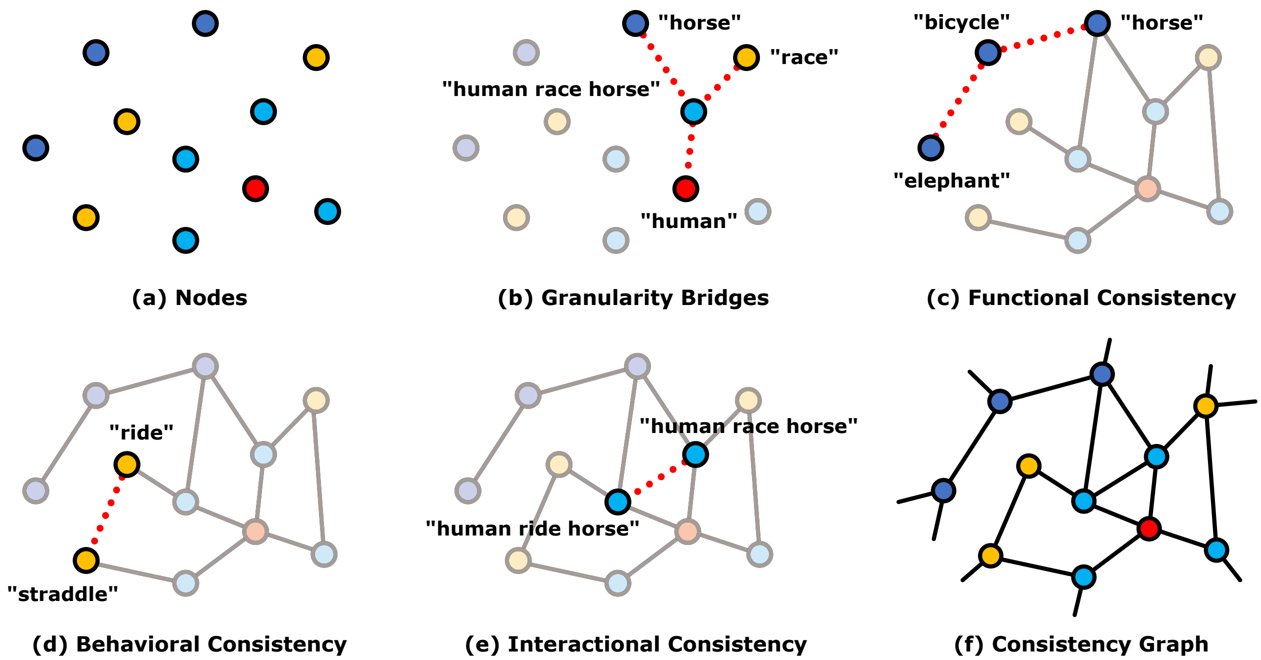


**Figure 5: Pipeline of constructing the consistency graph. a) Consistency graph contains human, object, action and interaction nodes. b) Each interaction node is linked with its entity nodes. c) Functional consistencies are represented by object-object connections. d) Behavioral consistencies are represented by action-action connections. e) Interactional consistencies are represented by interaction-interaction connections. f) Generalize the rules above and build consistency graph.**

*3.2.2 Fusion Block.* As described in Figure 4 (b), fusion block receives visual features of the human $a_h$, object $a_o$ and their spatial configuration $s$ as inputs, and does the same job as mapper blocks. The only difference is that dimensions of $a_h$, $a_o$ and $s$ are mapped to $1 \times 512$, $1 \times 512$ and $1 \times 256$ respectively using MLPs in advance. The concatenation of the mapped features serves as joint features of the entire human-object pair and be used to estimate $\varphi_k$ and $v_k$.

### 3.3 Semantic Embedding Network

To jointly capture multi-level consistencies among HOIs, we incorporate a knowledge graph, namely *consistency graph*, into the semantic embedding network, to help generate semantic embeddings of HOI categories in a transductive manner.

*3.3.1 Constructing the Graph.* Instead of using a large-scale knowledge graph, we distill the knowledge and construct a much smaller one, which only contains consistencies and compositional relations among HOIs and their entities. As illustrated in Figure 5, each HOI category refers to three entity nodes and one interaction node in the consistency graph. HOIs with shared entities would share the entity nodes as well. For instance, $\langle human, ride, bicycle\rangle$ and $\langle human, ride, horse\rangle$ are represented by four entity nodes "human", "ride", "bicycle" and "horse", as well as two interaction nodes "human ride bicycle" and "human ride horse".

We first add edges among interaction nodes and their corresponding entity nodes, which serve as bridges among different levels of consistencies. The other edges are defined based on the consistencies among objects, actions and interactions. That is, if two nodes are semantically consistent with each other, an edge would be added to enable message passing between them. We estimate the multi-level consistencies using cosine similarity by

$$\Theta_k(i,j) = \frac{z_k^i \cdot z_k^j}{\|z_k^i\|_2 \cdot \|z_k^j\|_2}, k \in \{a, o, t\} \tag{7}$$

where $k \in \{a, o, t\}$ indicates the type of a node, $\Theta_k(i,j)$ denotes the consistency between node $i$ and $j$, $z_k^i$ and $z_k^j$ represents visual-semantic joint features of the two nodes respectively. For each node, we link itself with only top $\varepsilon_k$ consistent nodes.

**Table 1: Role Detection results on V-COCO dataset under fully-supervised settings.**

| Method | Backbone | $mAP_{role}$ |
|---|---|---|
| Gupta *et al.* [13] | ResNet-50-FPN | 31.8 |
| InteractNet [11] | ResNet-50-FPN | 40.0 |
| GPNN [34] | DCN | 44.0 |
| iCAN [9] | ResNet-50 | 45.3 |
| TIN-$RP_{T2}C_D$ [25] | ResNet-50 | 48.7 |
| BAR-CNN [21] | Inception-ResNet | 43.6 |
| Wang *et al.* [41] | ResNet-50 | 47.3 |
| PMFNet [39] | ResNet-50 | 52.0 |
| VSGNet [37] | ResNet-152 | 51.8 |
| **ConsNet** (ours) | ResNet-50-FPN | **53.2** |

**Table 2: HOI Detection results on HICO-DET dataset under fully-supervised settings. *R* and *H* represent ResNet and Hourglass respectively.**

| Method | Backbone | Full | Rare | Non-Rare |
|---|---|---|---|---|
| Shen *et al.* [36] | VGG-19 | 6.46 | 4.24 | 7.12 |
| HO-RCNN [4] | CaffeNet | 7.81 | 5.37 | 8.54 |
| InteractNet [11] | R-50-FPN | 9.94 | 7.16 | 10.77 |
| GPNN [34] | DCN | 13.11 | 9.34 | 14.23 |
| iCAN [9] | R-50 | 14.84 | 10.45 | 16.15 |
| TIN-$RP_{T2}C_D$ [25] | R-50 | 17.22 | 13.51 | 18.32 |
| HOID [40] | R-50-FPN | 17.85 | 12.85 | 19.34 |
| Wang *et al.* [41] | R-50-FPN | 16.24 | 11.16 | 17.75 |
| Gupta *et al.* [14] | R-152 | 17.18 | 12.17 | 18.68 |
| PMFNet [39] | R-50-FPN | 17.46 | 15.65 | 18.00 |
| Peyre *et al.* [33] | R-50-FPN | 19.40 | 15.40 | 20.75 |
| VSGNet [37] | R-152 | 19.80 | 16.05 | 20.91 |
| **ConsNet** (ours) | R-50-FPN | **22.15** | **17.12** | **23.65** |
| Bansal *et al.* [1] | R-101 | 21.96 | 16.43 | 23.62 |
| PPDM [26] | H-104 | 21.73 | 13.78 | 24.10 |
| **ConsNet-F** (ours) | R-50-FPN | **24.39** | **17.10** | **26.56** |

**Table 3: HOI Detection results on HICO-DET dataset under zero-shot settings. *UC*, *UO* and *UA* denote unseen action-object combination, unseen object and unseen action scenarios respectively.**

| Method | Type | Full | Unseen | Seen |
|---|---|---|---|---|
| Shen *et al.* [36] | UC | 6.26 | 5.62 | - |
| Bansal *et al.* [1] | UC | 12.45±0.16 | 11.31±1.03 | 12.74±0.34 |
| **ConsNet** (ours) | UC | **14.48**±0.26 | **13.46**±1.24 | **14.74**±0.57 |
| Bansal *et al.* [1] | UO | 13.84 | 11.22 | 14.36 |
| **ConsNet** (ours) | UO | **14.48** | **13.51** | **14.67** |
| **ConsNet** (ours) | UA | **14.35** | **12.50** | **14.72** |

We propose a data-driven approach to generate the joint features of nodes. First, we collect all the visual features of humans and objects in HICO-DET dataset using a pre-trained object detector. These features are regarded as visual representations of actions and objects respectively. We then compute the average of all the visual representations with the same label to obtain the universal visual representations of these categories. Second, we adopt a pre-trained language model to generate word embeddings of node labels. Note that a label may contain multiple words, we fuse the word embeddings by computing their weighted sum. After collecting universal visual representations and word embeddings of node labels, we obtain the joint features of nodes by

$$z_k = (\rho_v \cdot \frac{q_k}{\|q_k\|_2}) \parallel (\rho_s \cdot \frac{e_k}{\|e_k\|_2}), k \in \{a, o, t\} \tag{8}$$

where $q_k$ and $e_k$ are visual and semantic representations of node labels, $\rho_v$ and $\rho_s$ are the weights of the representations above. The L-2 normalized, re-weighted and concatenated visual-semantic representations are then used to measure multi-level consistencies.

*3.3.2 Learning to Aggregate Semantic Representations.* Graph Attention Networks (GATs) [38] are first introduced for semi-supervised node classification. Instead of simply averaging the features of neighboring nodes like GCNs [20] or SGCs [43], GATs aggregate node features with a self-attention strategy. A single-level GAT layer can be represented as

$$h_i = \Big\Vert_{d=1}^{D} \tau(\sum_{j \in N_i} \mu_{i,j}^d \cdot \mathcal{W}^d \cdot h_j) \tag{9}$$

where $h_i$ and $h_j$ denotes the hidden states of node $i$ and $j$, $D$ indicates the number of attention heads, $\tau$ is the ReLU nonlinearity, $N_i$ represents the collection of node $i$ and its neighbours, $\mu_{i,j}^d$ is the attention coefficient learned by the model and $\mathcal{W}^d$ refers to the weight of this layer. In order to fix the output dimensions of the last GAT layer, we replace its concatenation with average operation. The attention coefficient $\mu_{i,j}^d$ can be predicted by

$$\mu_{i,j}^d = \frac{exp(\Gamma(\mathcal{W}^d \cdot h_i \parallel \mathcal{W}^d \cdot h_j))}{\sum_{k \in N_i} exp(\Gamma(\mathcal{W}^d \cdot h_i \parallel \mathcal{W}^d \cdot h_k))} \tag{10}$$

where $\mathcal{W}^d$ is the weight for estimating attention coefficient, $\Gamma$ is a single layer feed-forward network. The model uses a masked softmax to obtain the normalized attention coefficients $\mu_{i,j}^d$.

In this work, we adopt a three-layer GAT to propagate node features on the consistency graph. The input is a node feature matrix $\mathbf{Z} \in \mathbb{R}^{N \times C}$ given by a pre-trained ELMo [32]. After three layers of GATs, the node features are mapped to $D$ dimensions, which are the same with visual embeddings.

## 3.4 Model Learning

During training, visual embedding network learns to map visual features of human-object pairs into visual-semantic joint embedding space, while semantic embedding network learns to generate semantic embeddings of HOI categories. When testing, the semantic embeddings can be pre-computed and be used as templates of HOI categories. Since all the proposed components are differentiable, the whole model can be trained in an end-to-end manner. The overall objective of training is to minimize the distance among visual embeddings and semantic embeddings. We learn the parameters of the whole model by supervising $r_{h,o}$ and $\varphi_{h,o}$ with the following cross-entropy losses:

$$\mathcal{L}_i = -(u \cdot log(\varphi_{h,o}) + (1-u) \cdot log(1-\varphi_{h,o})) \tag{11}$$

$$\mathcal{L}_c = -\frac{1}{C}\sum_{k=1}^{C}(y_k \cdot log(r_{h,o}^k) + (1-y_k) \cdot log(1-r_{h,o}^k)) \tag{12}$$

where $u$ denotes interactiveness label and $y_k$ indicates HOI label. The interactiveness loss $\mathcal{L}_i$ and classification loss $\mathcal{L}_c$ are jointly optimized using their weighted sum by

$$\mathcal{L} = \mathcal{L}_i + \eta \cdot \mathcal{L}_c \tag{13}$$

where $\eta$ is a scale factor balancing the loss weights. Note that we optimize the classification loss only with positive samples and the interactiveness loss with both positive and negative samples.

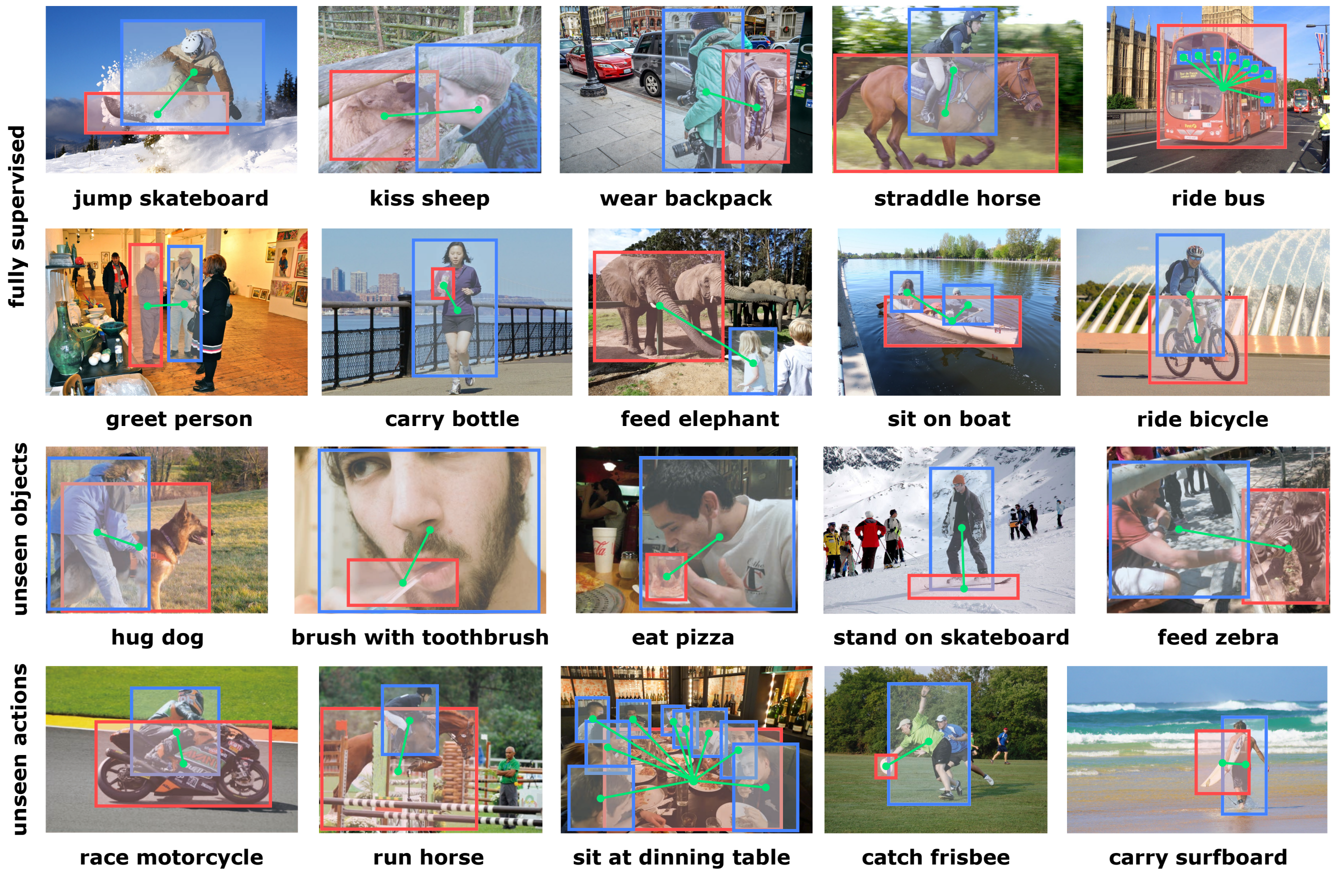


**Figure 6: Qualitative results on HICO-DET dataset. Our model has the ability to detect seen HOIs (the first two rows), HOIs with unseen objects (the third row) and HOIs with unseen actions (the last row). Note that non of the previous models can detect HOIs with unseen actions.**

## 4 EXPERIMENTS

In this section we evaluate the proposed method on the challenging V-COCO [13] and HICO-DET [4] datasets. We first evaluate our method under the fully-supervised settings on both V-COCO and HICO-DET datasets, following by unseen action-object combination, unseen object, and unseen action scenarios on HICO-DET dataset. Note that non of the previous methods have the ability to detect HOIs with unseen actions. An extensive ablation study is also reported after evaluations.

### 4.1 Datasets and Evaluation Metrics

V-COCO is a subset of MS-COCO dataset [28], it has 2,533 images for training, 2,867 images for validation and 4,946 images for testing. Each person is annotated with binary labels for 26 action categories. HICO-DET is another large-scale HOI detection dataset that extends annotations of HICO [5] from image-level to instance-level. The *trainval* split has 38,118 images while the *test* split has 9,658 images. It contains 117 action classes for 80 object classes, resulting in 600 HOI categories.

We follow the standard evaluation metric introduced by Chao *et al.* [4] that uses mean average precision (mAP) to measure the detection performance. An HOI detection is considered as a true positive when both the bounding boxes of the human and object have intersection over union (IoU) with a ground truth greater than 0.5, and the predicted HOI label is correct.

### 4.2 Implementation Details

We adopt Faster R-CNN [35] with ResNet-50-FPN as object detector. The same backbone is also used for feature extraction. We train the object detector on MS-COCO train2017 split using mmdetection [6]. When training *ConsNet*, we consider all the detections with confidence greater than 0.1 and make use of both ground truths and detected candidate pairs. When testing, we only consider up to 10 humans with confidence greater than 0.5 and up to 20 objects with confidence greater than 0.1 per image.

We add batch normalization [18] and ReLU nonlinearity after all hidden layers. For zero-shot settings, we also add extra dropout layers with rate 0.5 to prevent overfitting. Parameters of the backbone and ELMo are frozen during both training and testing. Each training mini-batch contains 64 samples with the ratio of positive and negative samples 1 : 3. For all experiments, we use Stochastic Gradient Descent (SGD) optimizer with initial learning rate 0.01,

**Table 4: Ablation study results on HICO-DET dataset under fully-supervised settings. ∅ means predicting HOI labels using visual embedding network directly.**

| Type | Embedder | Depth | Full | Rare | Non-Rare |
|---|---|---|---|---|---|
| ∅ | - | - | 18.90 | 10.57 | 21.40 |
| MLP | ELMo | 3 | 19.01 | 11.82 | 21.15 |
| SGC | ELMo | 3 | 19.63 | 14.85 | 21.05 |
| GCN | ELMo | 3 | 20.15 | 15.12 | 21.66 |
| SAGE | ELMo | 3 | 20.07 | 15.05 | 21.58 |
| GAT | ELMo | 2 | 21.16 | 16.82 | 22.46 |
| **GAT** | **ELMo** | **3** | **22.15** | **17.12** | **23.65** |
| GAT | ELMo | 4 | 21.12 | 16.35 | 22.54 |
| GAT | Word2Vec | 3 | 20.59 | 15.94 | 21.98 |
| GAT | GloVe | 3 | 20.63 | 15.66 | 22.12 |
| GAT | FastText | 3 | 20.58 | 15.68 | 22.04 |

momentum 0.9 and weight decay 0.0001. We drop the learning rate by 1/10 at epoch 3 and 4, and stop training at epoch 10.

## 4.3 Fully-Supervised HOI Detection

We first evaluate our model under fully-supervised settings. For both datasets, we train the model on *trainval* split and evaluate it on *test* split. The comparisons on V-COCO and HICO-DET datasets are shown in Table 1 and Table 2. Our method significantly outperforms the previous best models on each subset. Note that for HICO-DET dataset, the object detectors in Bansal *et al.* [1] and PPDM [26] are trained on MS-COCO and finetuned on HICO-DET, which may provide more potential true positives and largely reduce false positives. To be directly comparable, we also report the performance of our model with a finetuned detector called *ConsNet-F*, indicating that our method still achieves higher performance.

## 4.4 Zero-Shot HOI Detection

Shen *et al.* [36] first introduced the concept of detecting HOIs with unseen combinations of seen objects and actions. Bansal *et al.* [1] extended the task to detecting HOIs with unseen objects. We now extend the task further and introduce the scenario of detecting HOIs with unseen actions, which means the model should have the ability to analogize semantic representations of new actions based on similar actions or interactions, which is much more challenging than the two scenarios above. Below we report the performance comparisons under these scenarios on HICO-DET dataset.

*4.4.1 Unseen Combination Scenario.* The first three rows in Table 3 shows a comparison of our method with others under unseen combination scenario. We use the same 5 sets of 120 unseen classes as Bansal *et al.* and report the mean of the results. The comparison shows that our approach does much better on detecting unseen HOIs with seen objects and actions.

*4.4.2 Unseen Object Scenario.* Line 4 ~ 5 in Table 3 shows a comparison of our method with others under unseen object scenario. Our model outperforms over 2 mAPs than the previous best method on unseen classes while having similar performance on seen classes, indicating that our method can generalize to unseen objects better.

*4.4.3 Unseen Action Scenario.* In this scenario, we randomly select 22 actions, define them as *unseen* and remove all the training samples containing these actions. The full list of unseen action labels will be publicly available. We then train the model on the remaining samples and evaluate on the full *test* split. The last row in Table 3 reports the performance of our approach on detecting HOIs with unseen actions. The results show that our model has the ability to detect HOIs even if the action is previously unseen, which is quite challenging because transferring the knowledge of actions is much harder than objects. Moreover, our approach can even do slightly better than some early methods under fully-supervised settings.

*4.4.4 Quantitative Results.* Figure 6 shows quantitative results of both fully-supervised and zero-shot HOI detection using our method. Even if our model has never seen the objects or actions before, the semantic embedding network can still benefit from seen HOIs and generate semantic representations of unseen HOIs.

## 4.5 Ablation Study

We analyze the significance of the proposed knowledge-aware strategy for generating semantic representations by changing different types and depths of semantic embedding networks. All the experiments are performed on HICO-DET dataset under fully-supervised settings and the results are presented in Table 4.

Compared with not using semantic embedding network and simply using an MLP, HOI detection results on rare classes are largely improved with the use of GNNs. This is because the aggregation functions of GNNs can help transfer knowledge from non-rare classes to rare ones. The comparison also shows that with learnable attention coefficients, GATs are more flexible for generating semantic embeddings. Besides, the number of GAT layers matters. Deeper GATs can bring more learnable parameters, while it may cause the over-smoothing problem [23]. Performances are also considerably improved by changing word embeddings from Word2Vec [30], GloVe [31] or FastText [19] to ELMo [32]. The reason is that ELMo can better capture information at trigram level since the triplet is considered jointly as a whole.

# 5 CONCLUSION

In this work, we propose an end-to-end trainable framework for knowledge-aware human-object interaction detection by incorporating a consistency graph and exploiting GATs to propagate knowledge among nodes. Leveraging such a graph structure and message passing strategy, the model can capture and transfer knowledge about HOIs at different granularities and better generate semantic representations for rare or previously unseen HOIs.


# ACKNOWLEDGMENTS

This research is supported in part by Key-Area Research and Development Program of Guangdong Province, China with Grant 2019B010155002, National Natural Science Foundation of China Grant 91538203, US NSF Grant 1405594, and start-up funds from University at Buffalo.

## REFERENCES


[1] Ankan Bansal, Sai Saketh Rambhatla, Abhinav Shrivastava, and Rama Chellappa. 2020. Detecting Human-Object Interactions via Functional Generalization. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI).*

[2] Soravit Changpinyo, Wei-Lun Chao, Boqing Gong, and Fei Sha. 2016. Synthesized Classifiers for Zero-Shot Learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).* 5327–5336.

[3] Soravit Changpinyo, Wei-Lun Chao, and Fei Sha. 2017. Predicting Visual Exemplars of Unseen Classes for Zero-Shot Learning. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV).* 3476–3485.

[4] Yu-Wei Chao, Yunfan Liu, Xieyang Liu, Huayi Zeng, and Jia Deng. 2018. Learning to Detect Human-Object Interactions. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV).* 381–389.

[5] Yu-Wei Chao, Zhan Wang, Yugeng He, Jiaxuan Wang, and Jia Deng. 2015. HICO: A Benchmark for Recognizing Human-Object Interactions in Images. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV).* 1017–1025.

[6] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, Zheng Zhang, Dazhi Cheng, Chenchen Zhu, Tianheng Cheng, Qijie Zhao, Buyu Li, Xin Lu, Rui Zhu, Yue Wu, Jifeng Dai, Jingdong Wang, Jianping Shi, Wanli Ouyang, Chen Change Loy, and Dahua Lin. 2019. *MMDetection: Open MMLab Detection Toolbox and Benchmark.* Technical Report arXiv:1906.07155.

[7] Xinlei Chen, Abhinav Shrivastava, and Abhinav Gupta. 2013. NEIL: Extracting Visual Knowledge from Web Data. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV).* 1409–1416.

[8] Jia Deng, Nan Ding, Yangqing Jia, Andrea Frome, Kevin Murphy, Samy Bengio, Yuan Li, Hartmut Neven, and Hartwig Adam. 2014. Large-Scale Object Classification Using Label Relation Graphs. In *Proceedings of the European Conference on Computer Vision (ECCV).* 48–64.

[9] Chen Gao, Yuliang Zou, and Jia-Bin Huang. 2018. iCAN: Instance-Centric Attention Network for Human-Object Interaction Detection. In *Proceedings of the British Machine Vision Conference (BMVC).*

[10] Justin Gilmer, Samuel S Schoenholz, Patrick F Riley, Oriol Vinyals, and George E Dahl. 2017. Neural message passing for Quantum chemistry. In *Proceedings of the International Conference on Machine Learning (ICML).* 1263–1272.

[11] Georgia Gkioxari, Ross Girshick, Piotr Dollár, and Kaiming He. 2018. Detecting and Recognizing Human-Object Interactions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).* 8359–8367.

[12] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).* 6904–6913.

[13] Saurabh Gupta and Jitendra Malik. 2015. *Visual Semantic Role Labeling.* Technical Report arXiv:1505.04474.

[14] Tanmay Gupta, Alexander Schwing, and Derek Hoiem. 2019. No-Frills Human-Object Interaction Detection: Factorization, Layout Encodings, and Training Techniques. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV).* 9677–9685.

[15] Will Hamilton, Zhitao Ying, and Jure Leskovec. 2017. Inductive Representation Learning on Large Graphs. In *Advances in Neural Information Processing Systems (NeurlPS).* 1024–1034.

[16] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. 2017. Mask R-CNN. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV).* 2961–2969.

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).* 770–778.

[18] Sergey Ioffe and Christian Szegedy. 2015. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In *Proceedings of the International Conference on Machine Learning (ICML).*

[19] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. *Bag of Tricks for Efficient Text Classification.* Technical Report arXiv:1607.01759.

[20] Thomas N. Kipf and Max Welling. 2017. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations (ICLR).*

[21] Alexander Kolesnikov, Alina Kuznetsova, Christoph Lampert, and Vittorio Ferrari. 2019. Detecting Visual Relationships Using Box Attention. In *Proceedings of the IEEE International Conference on Computer Vision Workshops (ICCVW).*

[22] Yann LeCun, Lèon Bottou, Yoshua Bengio, and Patrick Haffner. 1998. Gradient-based learning applied to document recognition. *Proc. IEEE* 86, 11 (1998), 2278–2324.

[23] Guohao Li, Matthias Muller, Ali Thabet, and Bernard Ghanem. 2019. DeepGCNs: Can GCNs Go As Deep As CNNs?. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV).* 9267–9276.

[24] Yikang Li, Wanli Ouyang, Bolei Zhou, Kun Wang, and Xiaogang Wang. 2017. Scene Graph Generation From Objects, Phrases and Region Captions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).* 1261–1270.

[25] Yong-Lu Li, Siyuan Zhou, Xijie Huang, Liang Xu, Ze Ma, Hao-Shu Fang, Yanfeng Wang, and Cewu Lu. 2019. Transferable Interactiveness Knowledge for Human-Object Interaction Detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).* 3585–3594.

[26] Yue Liao, Si Liu, Fei Wang, Yanjie Chen, Chen Qian, and Jiashi Feng. 2020. PPDM: Parallel Point Detection and Matching for Real-Time Human-Object Interaction Detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).* 482–490.

[27] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. 2017. Feature Pyramid Networks for Object Detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).* 2117–2125.

[28] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft COCO: Common Objects in Context. In *Proceedings of the European Conference on Computer Vision (ECCV).* 740–755.

[29] Yao Lu. 2016. Unsupervised Learning on Neural Network Outputs: with Application in Zero-shot Learning. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI).*

[30] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. In *Advances in Neural Information Processing Systems (NeurlPS).*

[31] Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. GloVe: Global Vectors for Word Representation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP).* 1532–1543.

[32] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep Contextualized Word Representations. In *The North American Chapter of the Association for Computational Linguistics (NAACL).*

[33] Julia Peyre, Ivan Laptev, Cordelia Schmid, and Josef Sivic. 2019. Detecting Unseen Visual Relations Using Analogies. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV).* 1981–1990.

[34] Siyuan Qi, Wenguan Wang, Baoxiong Jia, Jianbing Shen, and Song-Chun Zhu. 2018. Learning Human-Object Interactions by Graph Parsing Neural Networks. In *Proceedings of the European Conference on Computer Vision (ECCV).* 401–417.

[35] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In *Advances in Neural Information Processing Systems (NeurlPS).* 91–99.

[36] Liyue Shen, Serena Yeung, Judy Hoffman, Greg Mori, and Li Fei-Fei. 2018. Scaling Human-Object Interaction Recognition Through Zero-Shot Learning. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV).* 1568–1576.

[37] Oytun Ulutan, A. S. M. Iftekhar, and Bangalore S. Manjunath. 2020. VSGNet: Spatial Attention Network for Detecting Human Object Interactions Using Graph Convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).* 13617–13626.

[38] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. Graph Attention Networks. In *International Conference on Learning Representations (ICLR).*

[39] Bo Wan, Desen Zhou, Yongfei Liu, Rongjie Li, and Xuming He. 2019. Pose-Aware Multi-Level Feature Network for Human Object Interaction Detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV).*

[40] Suchen Wang, Kim-Hui Yap, Junsong Yuan, and Yap-Peng Tan. 2020. Discovering Human Interactions With Novel Objects via Zero-Shot Learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).* 11652–11661.

[41] Tiancai Wang, Rao Muhammad Anwer, Muhammad Haris Khan, Fahad Shahbaz Khan, Yanwei Pang, Ling Shao, and Jorma Laaksonen. 2019. Deep Contextual Attention for Human-Object Interaction Detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV).* 5694–5702.

[42] Wei Wang, Vincent W. Zheng, Han Yu, and Chunyan Miao. 2019. A Survey of Zero-Shot Learning: Settings, Methods, and Applications. *ACM Transactions on Intelligent Systems and Technology (TIST)* 10, 2 (2019), 1–37.

[43] Felix Wu, Tianyi Zhang, Amauri Holanda de Souza Jr, Christopher Fifty, Tao Yu, and Kilian Q. Weinberger. 2019. Simplifying Graph Convolutional Networks. In *Proceedings of the International Conference on Machine Learning (ICML).*

[44] Danfei Xu, Yuke Zhu, Christopher B Choy, and Li Fei-Fei. 2017. Scene Graph Generation by Iterative Message Passing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).* 5410–5419.

[45] Ziming Zhang and Venkatesh Saligrama. 2016. Zero-Shot Learning via Joint Latent Similarity Embedding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).* 6034–6042.

[46] Jie Zhou, Ganqu Cui, Zhengyan Zhang, Cheng Yang, Zhiyuan Liu, Lifeng Wang, Changcheng Li, and Maosong Sun. 2018. *Graph Neural Networks: A Review of Methods and Applications.* Technical Report arXiv:1812.08434.