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

Ye Liu¹, Junsong Yuan², Chang Wen Chen²,³,⁴
¹Department of Resource and Environmental Sciences, Wuhan University, China
²Department of Computer Science and Engineering, State University of New York at Buffalo, USA
³Peng Cheng Laboratory, China
⁴School of Science and Engineering, The Chinese University of Hong Kong, Shenzhen, China
ye-liu@whu.edu.cn, {jsyuan,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 (human, action, object) 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. Code is available at https://github.com/yeliudev/ConsNet.

CCS CONCEPTS
• Computing methodologies → Activity recognition and understanding; Scene understanding.

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

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 (human, feed, cat) 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 [25], image retrieval [45], and visual question answering [13].

Most existing works on HOI detection [9, 11, 15, 26, 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 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, 26, 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, 15, 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. \( \langle \text{human, ride, elephant} \rangle \), because they can make use of their common sense to imagine what it would be like based on similar HOIs such as \( \langle \text{human, ride, bicycle} \rangle \) and \( \langle \text{human, feed, elephant} \rangle \), 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 propose \textit{functional consistency} which depicts the functional similarities among objects, and \textit{behavioral consistency} that represents the similarities of human behaviors when performing different actions. At trigram level, we present \textit{interactional consistency}, which denotes the holistic similarities among HOIs. We further construct an undirected graph, namely \textit{consistency graph}, to explicitly encode these relations. Each node in the consistency graph represents an HOI label or one of its entities. The three types of consistencies are encoded as edges among the nodes. That is, two object, 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 the HOI \( \langle \text{human, ride, bicycle} \rangle \) 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 to 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 [14]. Most previous works can be divided into compositional methods [1, 11, 15, 36] and non-compositional methods [4, 9, 26, 34, 39, 41]. Compositional methods learn separate detectors for objects and actions, then fuse the confidences to generate HOI detection results. However, these approaches suffer from the polysemy of action labels. Non-compositional methods avoid this problem by predicting fine-grained HOI labels directly, but they are restricted by the long-tail distribution of HOI categories. Recently introduced hybrid model [33] has shown that using multi-granularity representations of HOIs may solve the above contradiction. Nonetheless, all these
methods ignore the implicit relations among HOIs, thus we extend the hybrid model by incorporating 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 functions. Kipf et al. [21] extend the convolution operation [23] from euclidean data to non-euclidean data and proposed Graph Convolutional Networks (GCNs). Wu et al. [44] introduced SGCs to simplify GCNs by removing the non-linearities and merging the weights. Hamilton et al. [16] proposed GraphSAGE to realize inductive learning on graphs. In this work, we exploit Graph Attention Networks (GATs) [38] that incorporate multi-head attention mechanism to model the relations of neighboring nodes. The learned attention coefficients in GATs serve as the weights of consistencies.

**Zero-Shot Learning.** Most recent zero-shot learning methods can be divided into two protocols [43]. One is to learn semantic representations of categories that can be mapped to visual classifiers [2, 3]. The other is to make use of knowledge graphs to distill the knowledge [7, 8]. 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 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 between visual and semantic embeddings.

#### 3.1 Overview

Given an image \( x \) and a set of HOI categories of interest \( \mathcal{H} = \{1, \ldots, C\} \), 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 \( \mathcal{T} = \{(b_h, o, y_{ho})\} \), where \( b_h, b_o \in \mathbb{R}^d \) denotes bounding boxes of the human and the object, and \( y_{ho} \) represents a vector where \( y_{ho}^i \in \{0, 1\} \) 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 ones and classifying the remainders into multiple interaction categories. In the first stage, a pre-trained object detector is used to collect bounding boxes of humans \( b_h \) and objects \( b_o \), along with their corresponding detection confidences \( C_h, C_o \). We only keep top \( N_h \) detections with confidences \( C_h \) higher than a threshold \( \theta_h \), where \( k \in \{h, o\} \) denotes human or object. The candidates are then obtained by pairing up all the humans and objects extensively.

Recent works have shown that in most cases, the majority of humans and objects in an image are not interacting with each other.
Figure 4: Detailed architecture of visual embedding network. a) Mapper block takes visual features of human or object \(a_h, k \in \{h, o\}\) 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 \(\varphi_k\).

Such a severe imbalance between positive and negative candidates makes HOI classification challenging. To address this problem, Li et al. [26] proposed the strategy of non-active suppression (NIS) to filter out and suppress potential non-active candidates. In the second stage, we predict the class-irrelevant interactiveness \(\varphi_{h,o}\) for each candidate by

\[
\varphi_{h,o} = \sigma \left( \sum_{k \in \{h,o,a,t\}} \varphi_k \right)
\]

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

\[
P(y_i = 1 | x, b_h, b_o, c_h, c_o) = r_{h,o}^i \cdot \varphi_{h,o} \cdot c_h \cdot c_o
\]

where \(r_{h,o}^i\) 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-active or non-existent candidates. The HOI classification score \(r_{h,o}^i\) can be given by

\[
r_{h,o}^i = \sigma \left( \sum_{k \in \{h,o,a,t\}} \frac{v_k \cdot s_k^i}{\|v_k\| \cdot \|s_k^i\|} \cdot \gamma \right)
\]

where \(v_k\) denotes visual embeddings of the candidate, including human \(v_h\), object \(v_o\), action \(v_a\), and interaction \(v_i\). \(s_k^i\) 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 \(v_i\). 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 [18, 28], 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 [17] according to their bounding boxes. Spatial configuration of a candidate is computed by

\[
s = \left\| \sum_{k \in \{h,o\}} \left( x_k^i - dx_k \right) \left( y_k^i - dy_k \right) \right\| \psi
\]

where \(\psi\) denotes concatenation operation, \(x_k^i, y_k^i\) are coordinates of the human or object bounding box, \((dx_k, dy_k)\) and \(\psi\) represent the origin and area of the union box respectively. The computed spatial configuration \(s\) would be an \(1 \times 8\) 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 \(s\).

\[
P(v_m, v_m | x, b_h, b_o) = P(v_m, v_m | a_m), m \in \{h, o\}
\]

\[
P(v_n, v_n | x, b_h, b_o) = P(v_n, v_n | a_h, a_o, s), n \in \{a, t\}
\]

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 \(v_k\).
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 human-object pair and be used to estimate $\phi_k$ and $\psi_k$.

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.

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, $\langle \text{human, ride, bicycle}\rangle$ and $\langle \text{human, ride, horse}\rangle$ 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

$$
\Theta_k(i, j) = \frac{z^i_k \cdot z^j_k}{\|z^i_k\|_2 \cdot \|z^j_k\|_2}, \quad k \in \{a, o, t\}
$$

(7)

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

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

| Method          | Backbone          | mAP$_{\text{role}}$ |
|-----------------|-------------------|---------------------|
| Gupta et al. [14] | ResNet-50-FPN     | 31.8                |
| InteractNet [11] | ResNet-50-FPN     | 40.0                |
| GPNN [34]       | DCN               | 44.0                |
| iCAN [9]        | ResNet-50-FPN     | 45.3                |
| TIN-RP$_T$CD [26] | ResNet-50-FPN     | 48.7                |
| BAR-CNN [22]    | Inception-ResNet  | 43.6                |
| Wang et al. [41] | ResNet-50-FPN     | 47.3                |
| PMFNet [39]     | ResNet-50-FPN     | 52.0                |
| IP-Net [42]     | Hourglass-104     | 51.0                |
| VSGNet [37]     | ResNet-152        | 51.8                |
| **ConsNet (ours)** | ResNet-50-FPN     | **53.2**            |
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 the dataset using a pre-trained object detector. These features are regarded as visual representations of actions and objects respectively. Then, we compute the average of all the visual representations with the same label to obtain the universal visual representations of actions and objects. 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 binary cross-entropy losses:

\[
L_i = -(u \cdot \log(\varphi_{h,o}) + (1 - u) \cdot \log(1 - \varphi_{h,o}))
\]

\[
L_c = -\frac{1}{C} \sum_{k=1}^{C} (y_k \cdot \log(r_{h,o}^k) + (1 - y_k) \cdot \log(1 - r_{h,o}^k))
\]

where \( u \) denotes interactiveness label and \( y_k \) indicates HOI label. The interactiveness loss \( L_i \) and classification loss \( L_c \) are jointly optimized using their weighted sum by

\[
\mathcal{L} = L_i + \eta \cdot L_c
\]

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.

### 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_TOCD [26] | R-50-FPN | 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. [15] | R-152    | 17.18| 12.17| 18.68    |
| PMFNet [39]    | R-50-FPN | 17.46| 15.65| 18.00    |
| Peyret et al. [33] | R-50     | 19.40| 15.40| 20.75    |
| IP-Net [42]    | H-104    | 19.56| 12.79| 21.58    |
| VSGNet [37]    | R-152    | 19.80| 16.05| 20.91    |
| ConsNet-F (ours) | R-50-FPN | 22.15| 17.55| 23.52    |
| Bansal et al. [1]  | R-101    | 21.96| 16.43| 23.62    |
| PPDM [27]      | H-104    | 21.73| 13.78| 24.10    |
| ConsNet (ours) | R-50-FPN | 25.94| 19.35| 27.91    |

### 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  | Seen  | Unseen |
|--------------|------|-------|-------|--------|
| Shen et al. [36] | UC   | 6.26  | -     | 5.62   |
| Bansal et al. [1] | UC   | 12.45±0.16 | 12.74±0.34 | 11.31±1.03 |
| ConsNet (ours) | UC   | 19.81±0.32 | 20.51±0.62 | 16.99±1.67 |
| Bansal et al. [1] | UO   | 13.84 | 14.36 | 11.22 |
| ConsNet (ours) | UO   | 20.71 | 20.99 | 19.27 |
| ConsNet (ours) | UA   | 19.04 | 20.02 | 14.12 |
4 EXPERIMENTS

In this section, we evaluate the proposed method on the challenging V-COCO [14] and HICO-DET [4] datasets. We first evaluate our method under the fully-supervised settings on both of the datasets, following by zero-shot settings on HICO-DET dataset. The zero-shot settings includes three scenarios, i.e. unseen action-object combination, unseen object, and unseen action. An extensive ablation study is also reported after the evaluations.

4.1 Datasets and Evaluation Metrics

V-COCO is a subset of MS-COCO 2014 dataset [29], 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 of 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 the object detector. The same backbone and neck are also used for feature extraction. We train the object detector on MS-COCO 2017 dataset using MMDetection [6] and then freeze its weights. When training the HOI classifier, we use all the detections with confidence greater than 0.1 and make use of both ground truths and the 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 to reduce computational cost.

We add batch normalization [19] and ReLU nonlinearity [12] after all hidden layers. The classification losses of different samples are weighted to prevent overfitting. 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, momentum 0.9, and weight decay 0.0001. The linear warm-up policy starting from 0.001 learning rate for 500 iterations is adopted. All the models are trained for 5 epochs using cosine annealing learning rate schedule.
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 from non-rare classes to rare ones. The comparison also shows that with learnable attention coefficients, GATs are more flexible than other GNNs 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 [24], leading to a performance drop. Performances are also considerably improved by changing word embeddings from Word2Vec [30], GloVe [31], or FastText [20] to ELMo [32]. The probable reason is that ELMo can better capture information at trigram level since the triplet is considered jointly as a whole.

5 CONCLUSION

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 actions 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 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 Ablation Study

In order to analyze the significance of the proposed knowledge-aware strategy for generating semantic representations, we evaluate the models with different styles of semantic embedding networks and types of language models. All the experiments are performed on HICO-DET dataset under fully-supervised settings and the results are shown 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 than other GNNs 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 [24], leading to a performance drop. Performances are also considerably improved by changing word embeddings from Word2Vec [30], GloVe [31], or FastText [20] to ELMo [32]. The probable reason is that ELMo can better capture information at trigram level since the triplet is considered jointly as a whole.

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 |
|--------|----------|-------|--------|--------|----------|
| 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     | 3     | 22.16  | 16.82  | 22.46    |
| GAT    | ELMo     | 3     | 22.15  | 17.55  | 23.52    |
| 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    |

4.5 Qualitative Results

Figure 6 shows qualitative 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.

### 4.6 Conclusion

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.
