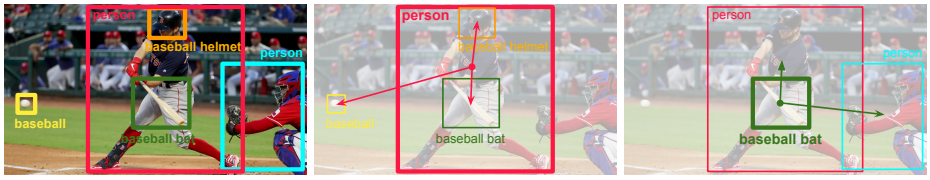
DRG: Dual Relation Graph for Human-Object Interaction Detection

Chen Gao, Jiarui Xu, Yuliang Zou, and Jia-Bin Huang

Virginia Tech
{chengao,jiaruixu,ylzou,jbhuang}@vt.edu



(a) Object detection (b) Human-centric (c) Object-centric

Fig. 1: **Human-object interaction (HOI) detection using dual relation graph.** Predicting each HOI in isolation is ambiguous due to the lack of context. In this work, we propose to leverage a dual relation graph. For each human node h, we obtain a *human-centric* subgraph where all object nodes are connected to h. Similarly, we can obtain an *object-centric* subgraph for each object node o. The human subgraph helps to adjust single HOI's prediction based on the same person's other HOIs. For example, knowing a **person** is wearing a **baseball helmet** and hitting a **baseball** suggests that the **person** may be holding a **baseball bat**. Similarly, the object subgraph helps to refine the HOI's prediction based on other HOIs associated with the same object. For example, knowing a **baseball bat** is held by a **person** lowers the chance that it is held by another **person**. Our method exploits such cues for improving HOI detection.

Abstract. We tackle the challenging problem of human-object interaction (HOI) detection. Existing methods either recognize the interaction of each human-object pair in isolation or perform joint inference based on complex appearance-based features. In this paper, we leverage an abstract spatial-semantic representation to describe each human-object pair and aggregate the contextual information of the scene via a dual relation graph (one *human-centric* and one *object-centric*). Our proposed dual relation graph effectively captures discriminative cues from the scene to resolve ambiguity from local predictions. Our model is conceptually simple and leads to favorable results compared to the state-of-the-art HOI detection algorithms on two large-scale benchmark datasets.

1 Introduction

Detecting individual persons and objects in isolation often does not provide sufficient information for understanding complex human activities. Moving beyond

detecting/recognizing individual objects, we aim to detect persons, objects, and recognize their interaction relationships (if any) in the scene. This task, known as human-object interaction (HOI) detection, can produce rich semantic information with visual grounding.

State-of-the-art HOI detection methods often use appearance features from the detected human/object instances as well as their relative spatial layout for predicting the interaction relationships [1, 3, 9, 12, 13, 14, 21, 24, 25, 32, 40, 41, 51]. These methods, however, often predict the interaction relationship between each human-object pair *in isolation*, thereby ignoring the contextual information in the scene. In light of this, several methods have been proposed to capture the contextual cues through iterative message passing [34, 43] or attentional graph convolutional networks [44]. However, existing approaches rely on complex appearance-based features to encode the human-object relation (e.g., deep features extracted from a union of two boxes) and do not exploit the informative spatial cues. In addition, the contexts are aggregated via a *densely connected* graph (where the nodes represent all the detected objects).

In this paper, we first propose to use spatial-semantic representation to describe each human-object pair. Specifically, our spatial-semantic representation encodes (1) the relative spatial layout between a person and an object and (2) the semantic word embedding of the object category. Using spatial-semantic representation for HOI prediction has two main advantages: First, it is invariant to complex appearance variations. Second, it enables knowledge transfer among object classes and helps with rare interaction during training and inference.

While such representations are informative, predicting HOI in isolation fails to leverage the contextual cues. In the example of Figure 1, a model might struggle to recognize that the person (in the red box) is hitting the baseball, by using only the spatial-semantic features from this particular human-object pair. Such ambiguity, however, may be alleviated if given the relation among different HOIs *from the same person*, e.g., this person is wearing a baseball helmet and holding a baseball bat. Similarly, we can exploit the relations among different HOIs *from the same object*. For example, a model may recognize both persons are holding the same baseball bat when making prediction independently. Knowing that the person (red box) is more likely to hold the baseball bat reduces the probability of another person (blue box) holding the same baseball bat. Inspired by these observations, we construct a *human-centric* and an *object-centric* HOI subgraph and apply attentional graph convolution to encode and aggregate the contextual information. We refer to our method as Dual Relation Graph (DRG).

Our contributions.

- We propose Dual Relation Graph, an effective method to capture and aggregate contextual cues for improving HOI predictions.
- We demonstrate that using the proposed spatial-semantic representation alone (without using appearance features) can achieve competitive performance compared to the state-of-the-art.
- We conduct extensive ablation study of our model, identifying contributions from individual components and exploring different model design choices.

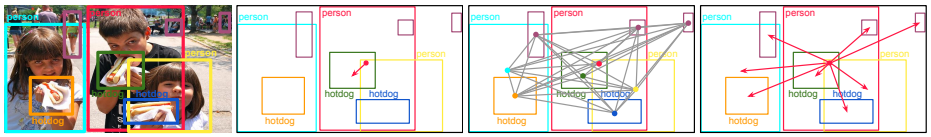


Fig. 2: **Leveraging contextual information.** Given object detections in the scene (a), existing HOI detection algorithms only perform *independent* prediction for each Human-object pair (b), ignoring the rich contextual cues. Recent methods in visual relationship detection (or scene graph generation) perform *joint inference* on a densely connected graph (c). While being general, the large number of relations among the dense connections makes the learning and inference on such a graph challenging. In contrast, our work leverages the human/object-centric graph to focus only on relevant contexts for improved HOI detection (d).

– We achieve competitive results compared with the state-of-the-art on the VCOCO and HICO-DET datasets.

2 Related Work

Human-object interaction detection. The task of human-object interaction detection aims to localize persons, object instances, as well as recognize the interactions (if any) between each pair of a person and an object. State-of-the-art HOI detection algorithms generally rely on two types of visual cues: (1) appearance features of the detected persons and objects (e.g., using the ROI pooling features extracted from a ConvNet) and (2) the spatial relationship between each human-object pair (e.g., using the bounding box transformation between the agent and the object [12, 13, 14], a two-channel interaction pattern [3, 9], or modeling the mutual contexts of human pose and object [14, 24, 46]). Recent advances focus on incorporating *contexts* to resolve potential ambiguity in interaction prediction based on independent human-object pairs, including pairwise body-parts [7, 40] or object-parts [51], instance-centric attention [9, 41], or message passing on a graph [34]. Our work shares similar spirits with these recent efforts as we also aim to capture contextual cues. The key difference lies in that the above approaches learn to aggregate contextual information from the other objects, body parts, or the scene background, while our method exploits *relations among different HOIs* to refine the predictions.

Inspired by the design of two-stage object detectors [35], recent works also show that filtering out candidate pairs with no relations using a relation proposal network [44] or an interactiveness network [24] improves the performance. Our method does not train an additional network for pruning unlikely relations. We believe that incorporating such a strategy may lead to further improvement.

Recent advances in HOI detection focus on tackling the long-tailed distributions of HOI classes. Examples include transferring knowledge from seen categories to unseen ones by an analogy transformation [31], performing data augmentation of semantically similar objects [1], or leveraging external knowledge graph [20].

While we do not explicitly address rare HOI classes, our method shows a small performance gap between rare and non-rare classes. We attribute this to the use of our abstract spatial-semantic representation.

Visual relationship detection. Many recent efforts have been devoted to detecting visual relationships [2, 5, 17, 22, 31, 50, 52]. Unlike object detection, the number of relationship classes can be prohibitively large due to the compositionality of object pairs, predicates, and limited data samples. To overcome this issue, some forms of language prior have been applied [27, 33]. Our focus in this work is on one particular class of relationship: human-centric interactions. Compared with other object classes, the possible interactions (the predicate) between a person and objects are significantly more fine-grained.

Scene graph. A scene graph is a graphical structure representation of an image where objects are represented as nodes, and the relationships between objects are represented as edges [30, 43, 44, 45, 49]. As the scene graph captures richer information beyond categorizing scene types or localizing object instances, it has been successfully applied to image retrieval [19], captioning [23], and generation [18]. Recent advances in scene graph generation leverage the idea of iterative message passing to capture contextual cues and produce a holistic interpretation of the scene [34, 43, 44, 49]. Our work also exploits contextual information but has the following distinctions: (1) Unlike existing methods that apply message passing to update *appearance features* (e.g., the appearance feature extracted from the union of human-object pair) at each step, we use an abstract spatial-semantic representation with an *explicit* encoding of relative spatial layout. (2) In contrast to prior works that use a single densely connected graph structure where edges connecting all possible object pairs, we operate on human-centric and object-centric subgraphs to focus on relevant contextual information specifically for HOI. Figure 2 highlights the differences between methods that capture contextual cues.

The mechanisms for dynamically capturing contextual cues for resolving ambiguity in local predictions have also been successfully applied to sequence prediction [38], object detection [16], action recognition [10, 37, 42], and HOI detection [9]. Our dual relation graph shares a similar high-level idea with these approaches but with a focus on exploiting the contexts of spatial-semantic representations.

Visual abstraction. The use of visual abstraction helps direct the focus to study the semantics of an image [53]. Visual abstraction has also been applied to learn common sense [39], forecasting object dynamics [8], and visual question answering [47]. Our work leverages the contexts of an abstract representation between human-object pairs for detecting HOIs.

Spatial-semantic representation. Spatial-semantic representation has also been applied in other problem domains such as image search [28], multi-class object detection [6], and image captioning [48].

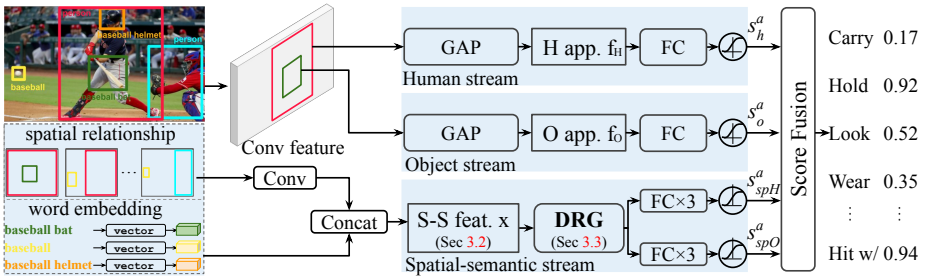


Fig. 3: **Overview of the proposed model.** Our network consists of three streams (human, object, and spatial-semantic). The human and object stream leverage the appearance feature $\mathbf{f}_h$ and $\mathbf{f}_o$. The spatial-semantic stream makes a prediction from the abstract spatial-semantic feature $\mathbf{x}$. We apply our proposed dual relation graph (DRG) to this stream. The three streams predict the scores s^a_h, s^a_o, s^a_{spH} and s^a_{spO}, which are fused to form final prediction.

3 Method

In this section, we present our network for HOI detection (Figure 3). We start with an overview of our network (Section 3.1). We then introduce the spatial-semantic representation (Section 3.2) and describe how we can leverage the proposed Dual Relation Graph (DRG) to propagate contextual information (Section 3.3). Finally, we outline our inference (Section 3.4) and the training procedure (Section 3.5).

3.1 Algorithm overview

Figure 3 provides a high-level overview of our HOI detection network. We decompose the HOI detection problem into two steps: (1) object detection and (2) HOI prediction. Following Gao et al. [9], we first apply an off-the-shelf object detector Faster R-CNN [35] to detect all the human/object instances in an image. We denote $\mathbb{H}$ as the set of human detections, and $\mathbb{O}$ as the set of object detections. Note that "person" is also an object category. We denote b_h as the detected bounding box for a person and b_o for an object instance. We use s_h and s_o to denote the confidence scores produced by the object detector for a detected person b_h and an object b_o, respectively. Given the detected bounding boxes b_h and b_o, we first extract the ROI pooled features and pass them into the human and object stream. We then pass the detected bounding boxes as well as the object category information to the spatial-semantic stream. We apply the proposed *Dual Relation Graph (DRG)* in the spatial-semantic stream. Lastly, we fuse the action scores from the three streams (human, object, and spatial-semantic) to produce our final predictions.

Human and object stream. Our human/object streams follow the standard object detection pipeline for feature extraction and classification. For each ROI pooled human/object feature, we pass it into a one-layer MLP followed by global average pooling and obtain the human appearance feature $\mathbf{f}_h$ and the object

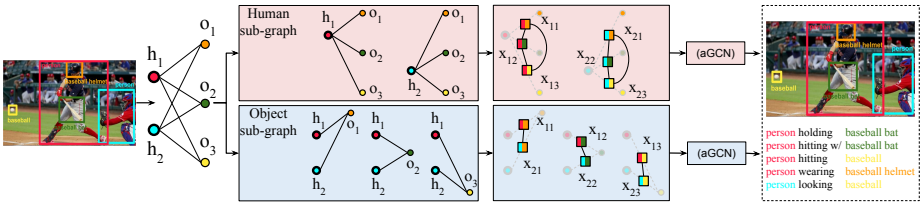


Fig. 4: **HOI detection using Dual Relation Graph.** (a) The input to our model are the detected objects in the given image. We denote $\mathbb{H}$ as the set of human detections, and $\mathbb{O}$ as the set of object detections. (b) We construct a *relation graph* from the detections where the two sets are $\mathbb{H}$ and $\mathbb{O}$. (c) For each human node h in $\mathbb{H}$, we obtain a human-centric sub-graph where all nodes in $\mathbb{O}$ are connected to h. Similarly, we can obtain an object-centric sub-graph for each object node o in $\mathbb{O}$. Note that "person" is also an object category. For simplicity, we do not show it in the figure. (d) In order to predict HOIs, we need to construct the HOI graph explicitly. Taking human sub-graph for example, we insert an HOI node x between human node h and object node o. We then connect all the HOI nodes and obtain the *human-centric* HOI sub-graph and the *object-centric* HOI sub-graph. (e) We iteratively update the HOI node feature via a trainable attentional graph convolutional network. This helps to aggregate the contextual information. (f) We fuse the scores from both sub-graphs and make the final HOI prediction.

appearance feature $\mathbf{f}_o$. We then apply a standard classification layer to obtain the A-dim action scores s_h^a (from human stream) and s_o^a (from object stream).

Spatial-semantic stream. Our inputs to this stream are the spatial-semantic features (described in Section 3.2). In an image, we pair all the detected persons in $\mathbb{H}$ with all the objects in $\mathbb{O}$, and extract spatial-semantic features for each human-object pair. We then pass all the features into our proposed Dual Relation Graph (Section 3.3) to aggregate the contextual information and produce updated features for each human-object pair. Our dual relation graph consists of a human-centric subgraph and an object-centric subgraph. These two subgraphs produce the action scores, s_{spH}^a and s_{spO}^a.

3.2 Spatial-semantic representation

We leverage the *abstract visual representation* of human-object pair for HOI prediction. The visual abstraction of a human-object pair allows us to construct representations that are invariant to intra-class appearance variations. In the context of human-object interaction, we consider the two important visual abstractions: (1) spatial relationship and (2) object category. [1]

[1] Other types of abstracted representation such as the pose of the person, the attribute of the person/object can also be incorporated into our formulation. We leave this to future work.

Capturing pairwise spatial relationship. Following [3, 9], we use the two-channel binary image representation to model the spatial relationship between a person and an object. To do so, we take the union of the two bounding boxes as a reference and rescale it to a fixed size. A binary image with two channels can then be created by filling value ones within the human bounding box in the first channel and filling value ones within the object bounding box in the second channel. The remaining locations are then filled with value 0. We then feed these two-channel binary images into a simple two-layer ConvNet to extract the spatial relation feature.

Capturing object semantics. We find that using spatial features by itself leads to poor results in predicting the interaction. To address this issue, we augment the spatial feature with the word embedding of each object's category, $vector(o)$, using fastText [29]. Let $\mathbf{x}_{ij}$ denote the *spatial-semantic feature* between the i-th person and the j-th object. We construct $\mathbf{x}_{ij} \in \mathbb{R}^{5708}$ by concatenating (1) the spatial feature and (2) the 300-dimensional word embedding vector.

3.3 Dual Relation Graph

Here, we introduce the Dual Relation Graph for aggregating the spatial-semantic features. Figure 4 illustrates the overall process.

Relation graph. Given the object detection results, i.e., instances in $\mathbb{H}$ and $\mathbb{O}$, we construct a relation graph (Figure 4b). There are two types of nodes in this graph, $\mathbb{H}$ (human) and $\mathbb{O}$ (object). For each node h in $\mathbb{H}$, we connect it to all the other nodes in $\mathbb{O}$. Similarly, for each node o in $\mathbb{O}$, we connect it to all nodes in $\mathbb{H}$.

Human-centric subgraph and object-centric subgraph. Unlike previous methods [34, 44], we do not use the densely connected graphs. To exploit the relation among different HOIs performed by *the same person*, we construct a *human-centric* subgraph. Similarly, we construct an *object-centric* subgraph for the HOIs performed on *the same object* (Figure 4c). So far, each node stands for an object instance detection. To explicitly represent the HOI, we insert an HOI node x_{ij} between each paired human node h_i and object node o_j. We then connect all the HOI nodes and obtain *human-centric* HOI subgraph and *object-centric* HOI subgraph (Figure 4d). We use the before mentioned *spatial-semantic feature* $\mathbf{x}_{ij}$ to encode each HOI node between the i-th person and the j-th object.

Contextual feature aggregation. With these two HOI subgraphs, we follow a similar procedure for propagating and aggregating features as in relation network [16], non-local neural network [42], and attentional graph convolutional network [44].

Human-centric HOI subgraph. To update node x_{ij}, we aggregate all the spatial-semantic feature of the nodes involving the same i-th person $\{x_{ij'}|j' \in \mathcal{N}(j)\}$. The feature aggregation can be written as:

$$\mathbf{x}_{ij}^{(l+1)} = \sigma\left(\mathbf{x}_{ij}^{(l)} + \sum_{j' \in \mathcal{N}(j)} \alpha_{jj'} W \mathbf{x}_{ij'}^{(l)}\right), \tag{1}$$

where $W \in \mathbb{R}^{5708\times5708}$ is a learned linear transformation that projects features into the embedding space, $\alpha_{ij'}$ is a learned attention weight.

We can rewrite this equation compactly with matrix operation:

$$\mathbf{x}_{ij}^{(l+1)} = \sigma\left(WX^{(l)}\boldsymbol{\alpha}_j\right) \tag{2}$$

$$u_{jj'} = \left(W_q\mathbf{x}_{ij'}^{(l)}\right)^\top \left(W_k\mathbf{x}_{ij}^{(l)}\right)/\sqrt{d_k} \tag{3}$$

$$\boldsymbol{\alpha}_j = \text{softmax}(\boldsymbol{u}_j), \tag{4}$$

where $W_q, W_k \in \mathbb{R}^{1024\times5708}$ are linear projections that project feature into a query and a key embedding space. Following [38], we calculate the attention weights using scaled dot-product, normalized by $\sqrt{d_k}$ where $d_k = 1024$ is the dimension of the key embedding space. We do not directly use the aggregated feature $\sigma\left(WX^{(l)}\boldsymbol{\alpha}_j\right)$ as our output updated feature. Instead, we add it back to the original spatial-semantic feature $\mathbf{x}_{ij}^{(l)}$. We then pass the addition through a LayerNorm to get the final aggregated feature on the human-centric subgraph.

$$\mathbf{x}_{ij}^{(l+1)} = \text{LayerNorm}\left(\mathbf{x}_{ij}^{(l)} + \sigma\left(WX^{(l)}\boldsymbol{\alpha}_j\right)\right). \tag{5}$$

The linear transformation W does not change the dimension of the input feature, thus the output $\mathbf{x}_{ij}^{(l+1)}$ has the same size as input $\mathbf{x}_{ij}^{(l)}$. As a result, we can perform several iterations of feature aggregation (Figure 4e). We explore the effectiveness of more iteration in Table 3(a).

Object-centric HOI subgraph. Similarly, to update node x_{ij}, we aggregate all the spatial-semantic feature of the nodes which involved the same object $\{x_{i'j}|i' \in \mathcal{N}(i)\}$. The two subgraphs have independent weights and aggregate contextual information independently.

3.4 Inference

For each human-object bounding box pair (b_h, b_o) in image I, we predict the score $S_{h,o}^a$ for each action $a \in \{1, \cdots, A\}$, where A denotes the total number of possible actions. The final score $S_{h,o}^a$ depends on (1) the confidence for the individual object detections (s_h and s_o), (2) the prediction score from the appearance of the person s_h^a and the object s_o^a, and (3) the prediction score based on the aggregated spatial-semantic feature, using human-centric and object-centric subgraph, s_{spH}^a and s_{spO}^a. We compute the HOI score $S_{h,o}^a$ for the human-object pair (b_h, b_o) as

$$S_{h,o}^a = s_h \cdot s_o \cdot s_h^a \cdot s_o^a \cdot s_{spH}^a \cdot s_{spO}^a \tag{6}$$

Note that we are not able to obtain the action scores using object s_o^a or the spatial-semantic stream for some classes of actions as they do not involve any objects (e.g., walk, smile). For those cases, we use only the score s_h^a from the human stream. For those actions, our final scores are $s_h \cdot s_h^a$.

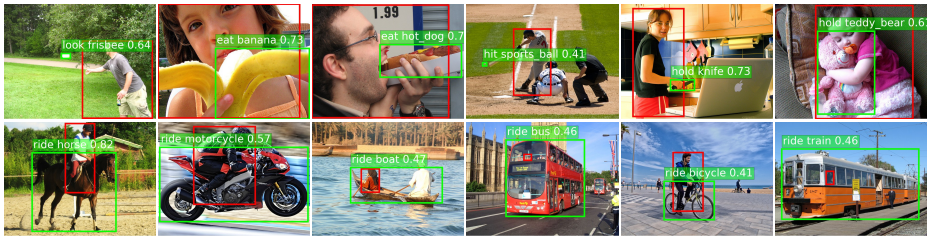


Fig. 5: **Sample HOI detections on V-COCO (first row) and HICO-DET (second row) *test* set.**

3.5 Training

HOI detection is a multi-label classification problem because a person can simultaneously perform different actions on different objects, e.g., sitting on a chair and reading a book. Thus, we minimize the cross-entropy loss *for each individual action class* between the ground-truth action label and the score produced from each stream. The total loss is the summation of the loss at each stream.

4 Experimental Results

In this section, we first outline our experimental setup, including datasets, metrics, and implementation details. We then report the quantitative results on two large-scale HOI benchmark datasets and compare the performance with the state-of-the-art HOI detection algorithms. Next, we show sample visual results on HOI detection. We conduct a detailed ablation study to quantify the contributions from individual components and validate our design choices. More results can be found in the supplementary material. We will make the source code and pre-trained models publicly available to foster future research.

4.1 Experimental setup

Datasets. V-COCO dataset [13] is constructed by augmenting the COCO dataset [26] with additional human-object interaction annotations. Each person is annotated with a binary label vector for 29 different action categories (five of them do not involve associated objects). **HICO-DET** [4] is a larger dataset containing 600 HOI categories over 80 object categories (same as [26]) with more than 150K annotated instances of human-object pairs. For applying our method on the HICO-DET dataset, we disentangle the 600 HOI categories into 117 object-agnostic action categories and train our network over these 117 action categories. At test time, we then combine the predicted action and the detected object and convert them back to the original 600 HOI classes. Note that the evaluation for the HICO-DET dataset remains the same.

Evaluation metrics. To evaluate the performance of our model, we adopt the commonly used role mean average precision (role mAP) [13] for both V-COCO and HICO datasets. The goal is to detect and correctly predict the ⟨

Table 1: **Comparison with the state-of-the-art on the V-COCO *test* set.** The best performance is in **bold** and the second best is underscored. Character * indicates that the method uses both VCOCO and HICO-DET training data. "S-S only" shows the performance of our spatial-semantic stream.

Method	Use human pose	Feature backbone	AP_{role}
VSRL [13]	-	ResNet-50-FPN	31.8
InteractNet [12]	-	ResNet-50-FPN	40.0
BAR-CNN [21]	-	Inception-ResNet	41.1
GPNN [34]	-	ResNet-101	44.0
iCAN [9]	-	ResNet-50	45.3
Wang et al. [41]	-	ResNet-50	47.3
RPNN [51]	✓	ResNet-50	47.5
$RP_{T2}C_D{}^\star$ [24]	✓	ResNet-50	48.7
PMFNet [40]	-	ResNet-50-FPN	48.6
PMFNet [40]	✓	ResNet-50-FPN	**52.0**
Ours (S-S only)	-	-	47.1
Ours	-	ResNet-50-FPN	51.0

`human, verb, object` $\rangle$ triplet. We consider a triplet as true positive if and only if it localizes the human and object accurately (i.e., with IoUs ≥ 0.5 w.r.t the ground truth annotations) and predicts the action correctly.

Implementation details. We build our network with the publicly available PyTorch framework. Following Gao et al. [9], we use the Detectron [11] with a feature backbone of ResNet-50 to generate human and object bounding boxes. For VCOCO, we conduct an ablation study on the validation split to determine the best threshold. We keep the detected human boxes with scores s_h higher than 0.8 and object boxes with scores s_o higher than 0.1. For HICO-DET, since there is no validation split available, we follow the setting in [32]. We use the score threshold 0.6 to filter out unreliable human boxes and threshold 0.4 to filter out unconfident object boxes. To augment the training data, we apply random spatial jitterring to the human and object bounding boxes and ensure that the IOU with the ground truth bounding box is greater than 0.7. We pair all the detected human and objects, and regard those who are not ground truth as negative training examples. We keep the negative to positive ratio to three.

We initialize our appearance feature backbone with the COCO pre-trained weight from Mask R-CNN [15]. We perform two iterations of feature aggregation on both *human-centric* and *object-centric* subgraphs. We train the three streams (human appearance, object appearance, and spatial-semantic) using the V-COCO *train* set. We use early stopping criteria by monitoring the validation loss. We train our network with a learning rate of 0.0025, a weight decay of 0.0001, and a momentum of 0.9 on both the V-COCO *train* set and HICO-DET *train* set. Training our network takes 14 hours on a single NVIDIA P100 GPU on V-COCO

Table 2: **Comparison with the state-of-the-art on HICO-DET *test* set.** The best performance is in **bold** and the second best is underscored. Character * indicates that the method uses both VCOCO and HICO-DET training data. For the object detector, "COCO" means that the detector is trained on COCO, while "HICO-DET" means that the detector is first pre-trained on COCO and then further fine-tuned on HICO-DET.

Method	Detector	Use human pose	Feature backbone	Default			Known Object		
				Full	Rare	Non Rare	Full	Rare	Non Rare
Shen et al. [36]	COCO	-	VGG-19	6.46	4.24	7.12	-	-	-
HO-RCNN [3]	COCO	-	CaffeNet	7.81	5.37	8.54	10.41	8.94	10.85
InteractNet [12]	COCO	-	ResNet-50-FPN	9.94	7.16	10.77	-	-	-
GPNN [34]	COCO	-	ResNet-101	13.11	9.34	14.23	-	-	-
iCAN [9]	COCO	-	ResNet-50	14.84	10.45	16.15	16.26	11.33	17.73
Wang et al. [41]	COCO	-	ResNet-50	16.24	11.16	17.75	17.73	12.78	19.21
Bansal et al. [1]	COCO	-	ResNet-101	16.96	11.73	18.52	-	-	-
RP_DC_D [24]	COCO	✓	ResNet-50	17.03	13.42	18.11	19.17	15.51	20.26
$RP_{T2}C_D{}^\star$ [24]	COCO	✓	ResNet-50	17.22	13.51	18.32	19.38	15.38	20.57
no-frills [14]	COCO	✓	ResNet-152	17.18	12.17	18.68	-	-	-
RPNN [51]	COCO	✓	ResNet-50	17.35	12.78	18.71	-	-	-
PMFNet [40]	COCO	-	ResNet-50-FPN	14.92	11.42	15.96	18.83	15.30	19.89
PMFNet [40]	COCO	✓	ResNet-50-FPN	17.46	15.65	18.00	20.34	17.47	21.20
Peyre et al. [32]	COCO	-	ResNet-50-FPN	**19.40**	14.63	**20.87**	-	-	-
Ours (S-S only)	COCO	-	-	12.45	9.84	13.23	15.77	12.76	16.66
Ours	COCO	-	ResNet-50-FPN	19.26	**17.74**	19.71	**23.40**	**21.75**	**23.89**
Bansal et al. [1]	HICO-DET	-	ResNet-101	21.96	16.43	23.62	-	-	-
Ours	HICO-DET	-	ResNet-50-FPN	**24.53**	**19.47**	**26.04**	27.98	23.11	29.43

and 24 hours on HICO-DET. At test time, our model runs at 3.3 fps for VCOCO and 5 fps for HICO-DET.

4.2 Quantitative evaluation

We report the main quantitative results in terms of AP_{role} on V-COCO in Table 1 and HICO-DET in Table 2. For the V-COCO dataset, our method compares favorably against state-of-the-art algorithms [24, 41, 51] except PMFNet [40], which uses human pose as an additional feature. Since pose estimation required additional training data (with pose annotations), we expect to see performance gain using human pose. PMFNet [40] also reports the AP_{role} *without* human pose, which is 2.4 mAP lower to our method. We also note that the spatial-semantic stream alone *without* using any visual features achieves a competitive performance (47.1 mAP) when compared with the state-of-the-art. This highlights the effectiveness of the abstract spatial-semantic representation and contextual information. Compared with methods that perform joint inference on densely connected graph [34], our approach produces significant performance gains.

For the HICO-DET dataset, our method also achieves competitive performance with state-of-the-art methods [14, 24, 31, 40]. Our method achieves the best performance for the *rare categories*, showing that our method handles the long-tailed distributions of HOI classes well.

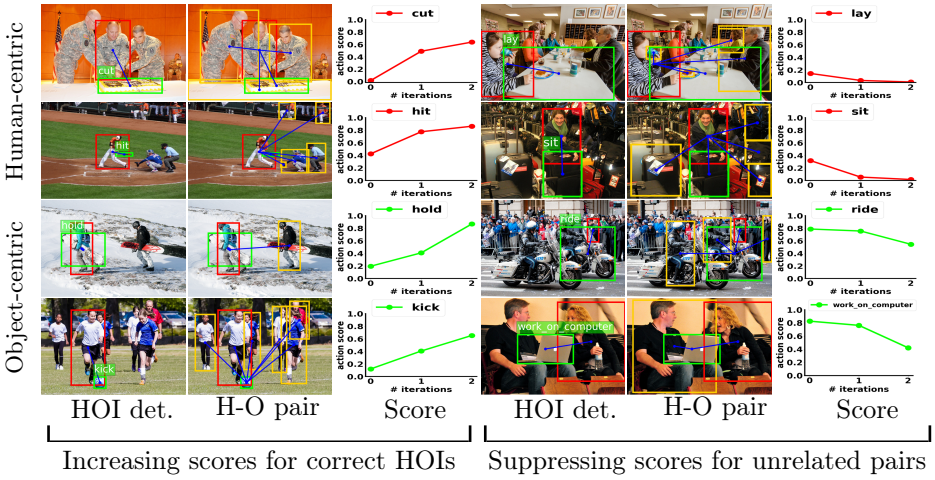


Fig. 6: **More iteration of feature aggregation leads to a more accurate prediction.** The human-centric and object-centric subgraph in the spatial-semantic stream propagates contextual information to produce increasingly accurate HOI predictions.

We note that the current best performing model [1] uses an object detector which is fine-tuned on HICO-DET *train* set using the annotated object bounding boxes. For a fair comparison, we also fine-tune our object detector on HICO-DET and report our result. Note that we do *not* re-train our model, but only replace the object detector at the test time.

Here, the large performance gain from fine-tuning the object detector may not reflect the progress on the HOI detection task. This is because the objects (and the associated HOIs) in the HICO-DET dataset are *not* fully annotated. Using a fine-tuned object detector can thus improve the HOI detection performance by exploiting such annotation biases.

4.3 Qualitative evaluation

HOI detection results. Here we show sample results on the V-COCO dataset and the HICO-DET dataset in Figure 5. We highlight the detected person and the associated object with red and green bounding boxes, respectively.

Visualizing the effect of the Dual Relation Graph. In Figure 6, we show the effectiveness of the proposed DRG. In the first two rows, we show that by aggregating contextual information, using the *human-centric* subgraph produces more accurate HOI predictions. Another example in the top right image indicates that the *human-centric* subgraph can also suppress the scores for unrelated human-object pairs. In the bottom two rows, we show four examples of how the *object-centric* subgraph propagates contextual information in each step to produce increasingly more accurate HOI predictions. For instance, the bottom

Table 3: **Ablation study on the V-COCO *val* set.** We show the role mAP AP_{role}.

(a) **More message passing iters.**

iter.	H graph	O graph	H + O
0-iter.	48.78	47.47	50.14
1-iter.	48.83	47.35	50.74
2-iter.	50.20	47.87	51.37

(b) **Feature used in DRG**

	mAP
App. feature (entire image)	35.69
App. feature (H-O union box)	46.93
Word2vec embedding	37.36
Spatial-semantic feature (ours)	51.37

(c) **Different subgraph**

H graph	O graph	mAP
-	-	50.14
✓	-	51.10
-	✓	50.78
✓	✓	51.37

(d) **Effectiveness of O subgraph**

	1-3	4-6	7+	all
H graph	57.89	52.77	50.96	51.10
H graph + O graph	58.28	53.75	51.06	51.37
Margin	+0.39	+0.98	+0.10	+0.27
% of testing images	68%	12%	20%	100%

right images show that for a person and an object without interaction, our model learns to suppress the predicted score by learning from the relationship of other HOI pairs associated with this particular object (laptop). In this example, the model starts with predicting a high score for the woman working on a computer. By learning from the relationship between the man and the computer, our model suppresses the score in each iteration.

4.4 Ablation study

We examine several design choices using the V-COCO *val* set.

More iteration of feature aggregation. Table 3(a) shows the performance using different iterations of feature aggregation. For either *human-centric* or *object-centric* subgraph, using more iterations of feature aggregation improves the overall performance. This highlights the advantages of exploiting contextual information among different HOIs. Performing feature aggregation on *both* subgraphs further improves the final performance.

Effectiveness of each subgraph. To validate the effectiveness of the proposed subgraph, we show different variants of our model in Table 3(c). Adding only *human-centric* subgraph improves upon the baseline model (without using any subgraph) by 0.96 absolute mAP, while adding only *object-centric* subgraph gives a 0.64 absolute mAP. More importantly, our results show that the performance gain of each subgraph is complementary to each other. To further validate the effectiveness of the *object-centric* subgraph, we show in Table 3(d) the breakdown of Table 3(c) in terms of the number of persons in the scene. The *object-centric* subgraph is less effective for cases with few people. For example, if there is only one person, the *object-centric* subgraph has no effect. For images with a moderate amount of persons (4-6), however, our *object-centric* subgraph shows a clear 0.98 mAP gain. As the number of persons getting larger (7+), the *object-centric*

Incorrect object Incorrect action

Fig. 7: **Failure cases of our method.**

subgraph shows a relatively smaller improvement due to clutter. Among the 2,867 testing images, 68% of them have only 1-3 persons. As a result, we do not see significant overall improvement.

Spatial-semantic representation. To demonstrate the advantage and effectiveness of the use of the abstract spatial-semantic representation, we show in Table 3(b) the comparison with alternative features, e.g., word2vec (as used in [27]) or appearance-based features. By using our spatial-semantic representation in the dual relation graph, we achieve 51.37 mAP. This shows a clear margin over the other alternative options, highlighting the contribution of spatial-semantic representation.

4.5 Limitations

While we demonstrated improved performance, our model is far from perfect. Below, we discuss two main limitations of our approach, with examples in Figure 7.

First, we leverage the off-the-shelf object detector to detect object instances in an image. The object detection does *not* benefit from the rich contextual cues captured by our method. We believe that a joint end-to-end training approach may help reduce this type of errors.

Second, our model may be confused by plausible spatial configuration and predicts incorrect action. In the third image, our model predicts that the person is sitting on a bench even though our model confidently predicts this person is standing and catching a Frisbee. Capturing the statistics of co-occurring actions may resolve such mistakes.

5 Conclusions

In this paper, we present a Dual Relation Graph network for HOI detection. Our core idea is to exploit the *global object layout* as contextual cues and use a *human-centric* as well as an *object-centric* subgraph to propagate and integrate rich relations among individual HOIs. We validate the efficacy of our approach on two large-scale HOI benchmark datasets and show our model achieves a sizable performance boost over the state-of-the-art algorithms. We also find that using the abstract spatial-semantic representation alone (i.e., without the appearance features extracted from a deep CNN) yields competitive accuracy, demonstrating a promising path of activity understanding through visual abstraction.

Acknowledgements We thank the support from Google Faculty Award.

References

1. Bansal, A., Rambhatla, S.S., Shrivastava, A., Chellappa, R.: Detecting human-object interactions via functional generalization. In: AAAI (2020)
2. Bilen, H., Vedaldi, A.: Weakly supervised deep detection networks. In: CVPR (2016)
3. Chao, Y.W., Liu, Y., Liu, X., Zeng, H., Deng, J.: Learning to detect human-object interactions. In: WACV (2017)
4. Chao, Y.W., Wang, Z., He, Y., Wang, J., Deng, J.: HICO: A benchmark for recognizing human-object interactions in images. In: CVPR (2015)
5. Dai, B., Zhang, Y., Lin, D.: Detecting visual relationships with deep relational networks. In: CVPR (2017)
6. Desai, C., Ramanan, D., Fowlkes, C.C.: Discriminative models for multi-class object layout. IJCV **95**(1), 1–12 (2011)
7. Fang, H.S., Cao, J., Tai, Y.W., Lu, C.: Pairwise body-part attention for recognizing human-object interactions. In: ECCV (2018)
8. Fouhey, D.F., Zitnick, C.L.: Predicting object dynamics in scenes. In: CVPR (2014)
9. Gao, C., Zou, Y., Huang, J.B.: iCAN: Instance-centric attention network for human-object interaction detection. In: BMVC (2018)
10. Girdhar, R., Carreira, J., Doersch, C., Zisserman, A.: Video action transformer network. In: CVPR (2019)
11. Girshick, R., Radosavovic, I., Gkioxari, G., Dollár, P., He, K.: Detectron. https://github.com/facebookresearch/detectron (2018)
12. Gkioxari, G., Girshick, R., Dollár, P., He, K.: Detecting and recognizing human-object interactions. In: CVPR (2018)
13. Gupta, S., Malik, J.: Visual semantic role labeling. arXiv preprint arXiv:1505.04474 (2015)
14. Gupta, T., Schwing, A., Hoiem, D.: No-frills human-object interaction detection: Factorization, appearance and layout encodings, and training techniques. In: ICCV (2019)
15. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: ICCV (2017)
16. Hu, H., Gu, J., Zhang, Z., Dai, J., Wei, Y.: Relation networks for object detection. In: CVPR (2018)
17. Hu, R., Rohrbach, M., Andreas, J., Darrell, T., Saenko, K.: Modeling relationships in referential expressions with compositional modular networks. In: CVPR (2017)
18. Johnson, J., Gupta, A., Fei-Fei, L.: Image generation from scene graphs. In: CVPR (2018)
19. Johnson, J., Krishna, R., Stark, M., Li, L.J., Shamma, D., Bernstein, M., Fei-Fei, L.: Image retrieval using scene graphs. In: CVPR (2015)
20. Kato, K., Li, Y., Gupta, A.: Compositional learning for human object interaction. In: ECCV (2018)
21. Kolesnikov, A., Lampert, C.H., Ferrari, V.: Detecting visual relationships using box attention. In: ICCV (2019)
22. Li, Y., Ouyang, W., Wang, X., Tang, X.: Vip-cnn: Visual phrase guided convolutional neural network. In: CVPR (2017)
23. Li, Y., Ouyang, W., Zhou, B., Wang, K., Wang, X.: Scene graph generation from objects, phrases and region captions. In: ICCV (2017)
24. Li, Y.L., Zhou, S., Huang, X., Xu, L., Ma, Z., Fang, H.S., Wang, Y.F., Lu, C.: Transferable interactiveness prior for human-object interaction detection. In: CVPR (2019)

25. Liao, Y., Liu, S., Wang, F., Chen, Y., Qian, C., Feng, J.: Ppdm: Parallel point detection and matching for real-time human-object interaction detection. In: CVPR (2020)
26. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: Common objects in context. In: ECCV (2014)
27. Lu, C., Krishna, R., Bernstein, M., Fei-Fei, L.: Visual relationship detection with language priors. In: ECCV (2016)
28. Mai, L., Jin, H., Lin, Z., Fang, C., Brandt, J., Liu, F.: Spatial-semantic image search by visual feature synthesis. In: CVPR (2017)
29. Mikolov, T., Grave, E., Bojanowski, P., Puhrsch, C., Joulin, A.: Advances in pre-training distributed word representations. In: LREC (2018)
30. Newell, A., Deng, J.: Pixels to graphs by associative embedding. In: NeurIPS (2017)
31. Peyre, J., Laptev, I., Schmid, C., Sivic, J.: Weakly-supervised learning of visual relations. In: ICCV (2017)
32. Peyre, J., Laptev, I., Schmid, C., Sivic, J.: Detecting rare visual relations using analogies. In: ICCV (2019)
33. Plummer, B.A., Mallya, A., Cervantes, C.M., Hockenmaier, J., Lazebnik, S.: Phrase localization and visual relationship detection with comprehensive linguistic cues. In: ICCV (2017)
34. Qi, S., Wang, W., Jia, B., Shen, J., Zhu, S.C.: Learning human-object interactions by graph parsing neural networks. In: ECCV (2018)
35. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. In: NeurIPS (2015)
36. Shen, L., Yeung, S., Hoffman, J., Mori, G., Fei-Fei, L.: Scaling human-object interaction recognition through zero-shot learning. In: WACV (2018)
37. Sun, C., Shrivastava, A., Vondrick, C., Murphy, K., Sukthankar, R., Schmid, C.: Actor-centric relation network. In: ECCV (2018)
38. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: NeurIPS (2017)
39. Vedantam, R., Lin, X., Batra, T., Lawrence Zitnick, C., Parikh, D.: Learning common sense through visual abstraction. In: ICCV (2015)
40. Wan, B., Zhou, D., Zhou, Y., Li, R., He, X.: Pose-aware multi-level feature network for human object interaction detection. In: ICCV (2019)
41. Wang, T., Anwer, R.M., Khan, M.H., Khan, F.S., Pang, Y., Shao, L., Laaksonen, J.: Deep contextual attention for human-object interaction detection. In: ICCV (2019)
42. Wang, X., Girshick, R., Gupta, A., He, K.: Non-local neural networks. In: CVPR (2018)
43. Xu, D., Zhu, Y., Choy, C.B., Fei-Fei, L.: Scene graph generation by iterative message passing. In: CVPR (2017)
44. Yang, J., Lu, J., Lee, S., Batra, D., Parikh, D.: Graph r-cnn for scene graph generation. In: ECCV (2018)
45. Yang, X., Zhang, H., Cai, J.: Shuffle-then-assemble: learning object-agnostic visual relationship features. In: ECCV (2018)
46. Yao, B., Fei-Fei, L.: Modeling mutual context of object and human pose in human-object interaction activities. In: CVPR (2010)
47. Yi, K., Wu, J., Gan, C., Torralba, A., Kohli, P., Tenenbaum, J.B.: Neural-symbolic vqa: Disentangling reasoning from vision and language understanding. In: NeurIPS (2018)
48. Yin, X., Ordonez, V.: Obj2text: Generating visually descriptive language from object layouts. In: EMNLP (2017)

49. Zellers, R., Yatskar, M., Thomson, S., Choi, Y.: Neural motifs: Scene graph parsing with global context. In: CVPR (2018)
50. Zhang, H., Kyaw, Z., Yu, J., Chang, S.F.: Ppr-fcn: Weakly supervised visual relation detection via parallel pairwise r-fcn. In: ICCV (2017)
51. Zhou, P., Chi, M.: Relation parsing neural network for human-object interaction detection. In: ICCV (2019)
52. Zhuang, B., Liu, L., Shen, C., Reid, I.: Towards context-aware interaction recognition for visual relationship detection. In: ICCV (2017)
53. Zitnick, C.L., Parikh, D.: Bringing semantics into focus using visual abstraction. In: CVPR (2013)