A Graph-based Interactive Reasoning for Human-Object Interaction Detection

Dongming Yang[1] and **Yuexian Zou**[1,2*]

[1]School of ECE, Peking University, Shenzhen, China
[2]Peng Cheng Laboratory, Shenzhen, China
{yangdongming, zouyx}@pku.edu.cn

Abstract

Human-Object Interaction (HOI) detection devotes to learn how humans interact with surrounding objects via inferring triplets of ⟨ human, verb, object ⟩. However, recent HOI detection methods mostly rely on additional annotations (e.g., human pose) and neglect powerful interactive reasoning beyond convolutions. In this paper, we present a novel graph-based interactive reasoning model called Interactive Graph (abbr. in-Graph) to infer HOIs, in which interactive semantics implied among visual targets are efficiently exploited. The proposed model consists of a project function that maps related targets from convolution space to a graph-based semantic space, a message passing process propagating semantics among all nodes and an update function transforming the reasoned nodes back to convolution space. Furthermore, we construct a new framework to assemble in-Graph models for detecting HOIs, namely in-GraphNet. Beyond inferring HOIs using instance features respectively, the framework dynamically parses pairwise interactive semantics among visual targets by integrating two-level in-Graphs, i.e., scene-wide and instance-wide in-Graphs. Our framework is end-to-end trainable and free from costly annotations like human pose. Extensive experiments show that our proposed framework outperforms existing HOI detection methods on both V-COCO and HICO-DET benchmarks and improves the baseline about 9.4% and 15% relatively, validating its efficacy in detecting HOIs.

1 Introduction

The task of Human-Object Interaction (HOI) detection aims to localize and classify triplets of ⟨ human, verb, object ⟩ from a still image. Beyond detecting and comprehending instances, e.g., object detection [Ren *et al.*, 2017; He *et al.*, 2017], segmentation [Yang *et al.*, 2018] and human pose estimation [Fang *et al.*, 2017], detecting HOIs requires a deeper
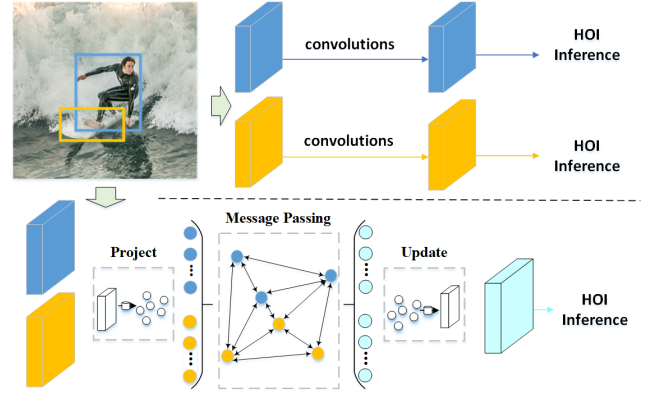
---

*Contact Author

Figure 1: A simplified schematic of how our framework facilitates HOI detection. The baseline framework (above) adopts convolutional features directly to infer HOIs and neglects interactive reasoning. We present a novel model (bottom) that reasons interactive semantics among visual targets to learn HOI-specific representation.

understanding of visual semantics to depict complex relationships between human-object pairs. HOI detection is related to action recognition [Sharma *et al.*, 2015] but presents different challenges, e.g., an individual can simultaneously take multiple interactions with surrounding objects. Besides, associating ever-changing roles with various objects leads to finer-grained and diverse samples of interactions.

Most existing HOI detection approaches infer HOIs by directly employing the appearance features of a person and an object extracted from Convolutional Neural Networks (CNNs) [He *et al.*, 2016], which may neglect the detailed and high-level interactive semantics implied between the related targets. To remedy the limitation above, a number of algorithms exploit additional contextual cues from the image such as human intention [Xu *et al.*, 2019] and attention boxes [Kolesnikov *et al.*, 2018]. More recently, several works have taken advantage of additional annotations, e.g., human pose [Xu *et al.*, 2019; Li *et al.*, 2019] and body parts [Zhou and Chi, 2019; Wan *et al.*, 2019]. Although incorporating contextual cues and annotations generally benefits feature expression, it brings several drawbacks.

Firstly, stack of convolutions and contextual cues are deficient in modelling HOIs since recognizing HOIs requires reasoning beyond feature extraction. However, dominant meth-

ods are limited by treating each visual component separately without considering crucial semantic dependencies among related targets (i.e., scene, human and object). Secondly, employing additional human pose and body parts in HOI detection algorithms brings a large computational burden.

To address these issues, we propose a novel graph-based model called Interactive Graph (abbr. in-Graph) to infer HOIs by reasoning and integrating strong interactive semantics among scene, human and object. As illustrated in Figure 1, our model goes beyond current approaches lacking the capability to reason interactive semantics. In particular, in-Graph model contains three core procedures, i.e., a project function, a message passing process and an update function. Here, the project function generates a unified space to make two related targets syncretic and interoperable. The message passing process further integrates semantic information by propagating messages among nodes. Finally, the update function transforms the reasoned nodes to convolution space, providing enhanced representation for HOI-specific modeling.

Based on the proposed in-Graph model, we then offer a general framework referred to as in-GraphNet to implicitly parse scene-wide interactive semantics and instance-wide interactive semantics for inferring HOIs rather than treat each visual target separately. Concretely, the proposed in-GraphNet is a multi-stream network assembling two-level in-Graph models (i.e., scene-wide in-Graph and instance-wide in-Graph). The final HOI predictions are made by combining all exploited semantics. Moreover, our framework is free from additional annotations such as human pose.

We perform extensive experiments on two public benchmarks, i.e., V-COCO [Yatskar *et al.*, 2016] dataset and HICO-DET [Chao *et al.*, 2018] dataset. Our method provides obvious performance gain compared with the baseline and outperforms the state-of-the-art methods (both pose-free and pose-based methods) by a sizable margin. We also provide detailed ablation studies of our method to facilitate the future research.

2 Related Work

2.1 Contextual Cues in HOI Detection

The early human activity recognition [Idrees *et al.*, 2017] task is confined to scenes containing single human-centric action and ignores spatial localization of the person and related object. Therefore, Gupta [Yatskar *et al.*, 2016] introduced visual semantic role labeling to learn interactions between human and object. HO-RCNN [Chao *et al.*, 2018] introduced a three-branch architecture with one branch each for a human candidate, an object candidate, and an interaction pattern encoding the spatial position of the human and object. Recently, several works have taken advantage of contextual cues and detailed annotations to improve HOI detection. Auxiliary boxes [Gkioxari *et al.*, 2015] were employed to encode context regions from the human bounding boxes. InteractNet [Gkioxari *et al.*, 2018] extended the object detector Faster R-CNN [Ren *et al.*, 2017] with an additional branch and estimated an action-specific density map to identify the locations of interacted objects. iHOI [Xu *et al.*, 2019] utilized human gaze to guide the attended contextual regions in a weakly-supervised setting for learning HOIs.

Very recently, human pose [Fang *et al.*, 2017] has been widely adopted as an additional cue to tackle HOI detection. Pair-wise human parts correlation [Fang *et al.*, 2018] was exploited to learn HOIs. TIN [Li *et al.*, 2019] combined human pose and spatial configuration to encode pose configuration maps. PMFNet [Wan *et al.*, 2019] developed a multi-branch network to learn a pose-augmented relation representation to incorporate interaction context, object features and detailed human parts. RPNN [Zhou and Chi, 2019] introduced an object-bodypart graph and a human-bodypart graph to capture relationships between body parts and surrounding instances (i.e., human and object).

Although extracting contextual evidence benefits feature expression, it is not favored since additional annotations and computation are indispensable. For example, pose-based approaches are inseparable from pre-trained human pose estimators like Mask R-CNN [He *et al.*, 2017] and AlphaPose [Fang *et al.*, 2017], which bring large workload and computational burden.

2.2 Semantic Reasoning

Attention mechanism [Sharma *et al.*, 2015] in action recognition helps to suppress irrelevant global information and highlight informative regions. Inspired by action recognition methods, iCAN [Gao *et al.*, 2018] exploited an instance-centric attention mechanism to enhance the information from regions and facilitate HOI classification. Furthermore, Contextual Attention [Wang *et al.*, 2019] proposed a deep contextual attention framework for HOI detection, in which context-aware appearance features for human and object were captured. PMFNet [Wan *et al.*, 2019] focused on pose-aware attention for HOI detection by employing human parts. Overall, methods employing attention mechanisms learn informative regions but treat each visual target (i.e., scene, human and object) separately, which are still insufficient to exploit interactive semantics for inferring HOIs.

Graph Parsing Neural Network (GPNN) [Qi *et al.*, 2018] introduced a learnable graph-based structure, in which HOIs were represented with a graph structure and parsed in an end-to-end manner. The above structure was a generalization of Message Passing Neural Network [Gilmer *et al.*, 2017] using a message function and a Gated Recurrent Unit (GRU) to iteratively update states. GPNN was innovative but showed some limitations. Firstly, it reasoned interactive features at the coarse instance-level (i.e., each instance was encoded as an infrangible node), which suffered from handling complex interactions. In addition, it required iteratively message passing and updating. Thirdly, it excluded semantic information from the scene in inferring HOIs. Lately, RPNN [Zhou and Chi, 2019] introduced a complicated structure with two graph-based modules incorporated together to infer HOIs, but in which fine-grained human pose was required as prior information.

In this paper, we aim to develop a novel graph-based model to provide interactive semantic reasoning between visual targets. Instead of coarse and iterative message passing between instances, our model captures pixel-level interactive semantics between targets all at once. Furthermore, our model is free from costly annotations like human pose.
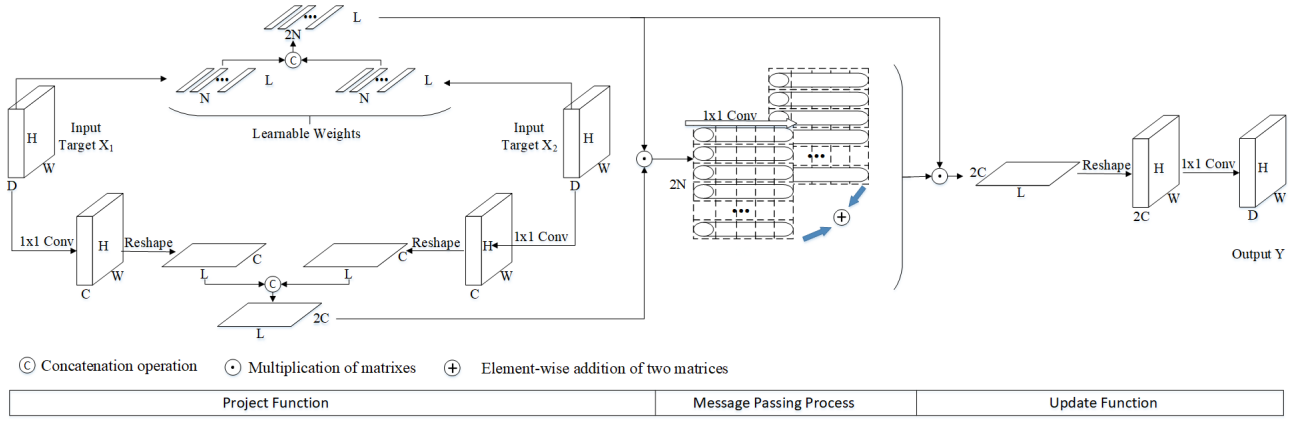
Figure 2: The detailed design of proposed in-Graph model. The model consists of three core procedures, which are a project function, a message passing process and an update function. The in-Graph takes convolutional features of two visual targets as inputs and reasons strong interactive semantics between them by a graph-based structure. Finally, reasoned semantics are output in same form as inputs.

3 Proposed Method

3.1 Overview of in-Graph

The detailed design of proposed in-Graph model is provided in Figure 2. Scene, human and object are three semantic elements been considered as three visual targets in our model, referred to as T_{scene}, T_{human}, and T_{object}, respectively. The proposed in-Graph takes two targets once to conduct pixel-level interactive reasoning. Each of the targets takes convolutional feature $X \in \Re^{H*W*D}$ according to the corresponding boxes (i.e., the whole image, human candidate boxes and object candidate boxes) as input. Here $H*W$ denotes locations and D denotes feature dimension.

We first propose a project function to map two feature tensors into a graph-based semantic space named interactive space, where a fully-connected graph structure can be built. Based on the graph structure, message passing process is then adopted as modeling the interaction among all nodes by propagating and aggregating interactive semantics. Finally, the update function provides a reversed projection over interactive space and output feature Y, enabling us to utilize the reasoned semantics in convolution space. We then describe its architecture in details and explain how we apply it into HOI detection task.

3.2 Project Function

Project function aims to provide an pattern $P(\Delta)$ to fuse two targets together, after which message passing process can be efficiently computed. The calculation process of $P(\Delta)$ can be divided into three parts: a feature conversion denoting as $\varphi(X_1, X_2; W_\varphi)$, a weights inference denoting as $\theta(X_1, X_2; W_\theta)$ and a linear combination, where W_φ and W_θ are learnable parameters. Finally, the function outputs a matrix $V = P(X_1, X_2) \in \Re^{2N*2C}$, where X_1 and X_2 are input feature tensors, $2N$ denotes the number of nodes in interactive space and $2C$ refers to dimension.

Feature conversion. Given feature tensors of two targets $X_1, X_2 \in \Re^{H*W*D}$, we first employ $1*1$ convolutions to reduce the dimensions of X_1 and X_2 to C, thus the computation of the block can be valid decreased. The obtained tensors

are then reshaped from $H*W*C$ to planar $L*C$, obtaining $X_1^r, X_2^r \in \Re^{L*C}$, where its two-dimensional location pixels of $H*W$ are converted to one-dimensional vector L. After that, a concatenation operation is adopted to integrate X_1^r and X_2^r by dimensional C, obtaining $X^r = [X_1^r, X_2^r] \in \Re^{L*2C}$.

Weights inference. Here, we infer learnable projection weights B so that semantic information from original features can be weighted aggregated. Instead of designing complicated calculations, we simply use convolution layers to generate the dynamically weights. In this step, X_1 and X_2 are feed into $1*1$ convolutions to obtain weight tensors with channel of N. Obtained feature tensors are then reshaped as planar $B_1, B_2 \in \Re^{N*L}$. Finally, integrated projection weights $B = [B_1, B_2] = [b_1, b_2, \ldots b_i, \ldots, b_{2N}] \in \Re^{2N*L}$ are obtained by a concatenation operation.

Linear combination. Since the project function involves two targets, linear combination is a necessary step to aggregate the semantic information and transform targets to the unified interactive space. In particular, node $v_i \in V$ in interactive space is calculated as follow. Here $x_j^r \in \Re^{1*2C}$, $v_i \in \Re^{1*2C}$.

$$v_i = \sum_{\forall j} b_{ij} x_j^r. \tag{1}$$

The proposed project function is simple and fast since all parameters are end-to-end learnable and come from $1*1$ convolutions. Such a function achieves semantic fusion between two targets and maps them into an interactive space effectively.

3.3 Message Passing and Update Function

Message Passing

After projecting targets from convolution space to interactive space, we have a structured representation of a fully-connected graph $G = (V, E, A)$, where each node contains a feature tensor as its state and all nodes are considered as fully-connected with each other. Based on the graph structure, message passing process is adopted to broadcast and integrate semantic information from all nodes over the graph.

GPNN [Qi *et al.*, 2018] applies an iterative process with GRU to enable nodes to communicate with each other, whereas it needs to run several times iteratively towards convergence. We reason interactive semantics over the graph structure by adopting a single-layer convolution to efficiently build communication among nodes. In our model, the message passing functions $M(\Delta)$ is computed by:

$$M(V) = \omega(AV) = Conv1D(V) \oplus V. \tag{2}$$

Here A denotes the adjacency matrix among nodes learned by gradient decent during training, reflecting the weights for edge E. $\omega(\Delta)$ denotes the state update of nodes. In our implementation, the operation of $Conv1D(V)$ is a channel-wise 1D convolution layer that performs Laplacian smoothing [Li *et al.*, 2018] and propagates semantic information among all connected nodes. After information diffusion, the $\oplus$ implements addition point to point which updates the hidden node states according to the incoming messages.

Update Function

To apply above reasoning results into convolutional network, an update function $U(\Delta)$ provides a reverse projection for reasoned nodes from interactive space to convolution space, which output Y as a new feature tensor. Given the reasoned nodes $V' \in \Re^{2N*2C}$, update function first adopts a linear combination as follows:

$$y_i = \sum_{\forall j} b_{ij} v'_j. \tag{3}$$

Where the projection weights B is transposed and reused, here $v'_j \in \Re^{1*2C}$, $y_i \in \Re^{1*2C}$.

After the linear combination, we reshape the obtained tensor from planar $L * 2C$ to three-dimensional $H * W * 2C$. Finally, a $1 * 1$ convolution is attached to expand the feature dimensions from C to D to match the inputs. In this way, updated features in convolution space can play its due role in the following schedule.

3.4 in-GraphNet

Assembling in-Graph Model

Our in-Graph model improves the ability of modelling HOIs by employing interactive semantic reasoning beyond stack of convolutions. It is noted that the human visual system is able to progressively capture interactive semantics from the scene and related instances to recognize a HOI. Taking the HOI triplet $\langle$ human, surf, surfboard $\rangle$ as an example, the scene-wide interactive semantics connected with the scene (e.g., sea) and instances (e.g., human, surfboard) can be captured as prior knowledge and instance-wide interactive semantics between the person and surfboard are learned to further recognize the verb (i.e., surf) and disambiguate other candidates (e.g., carry). Inspired by this human perception, we assign in-Graphs in two levels to build in-GraphNet, which are scene-wide level and instance-wide level. The scene-wide in-Graphs contain a human-scene in-Graph and an object-scene in-Graph, the instance-wide in-Graph refers to human-object in-Graph.

An overview of proposed in-GraphNet is shown in Figure 3. Since in-Graph models are light-weight, it can be easily incorporated into existing CNN architectures. ResNet-50
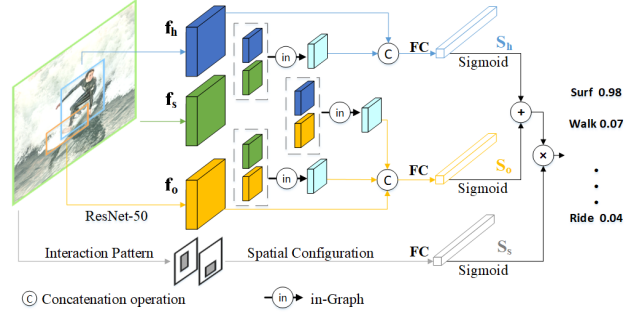

Figure 3: An overview of in-GraphNet. The framework consists of three branches, which are a human-centric branch, an object-centric branch and a spatial configuration branch. We integrate two scene-wide in-Graphs and an instance-wide in-Graph into in-GraphNet to infer HOIs.

[He *et al.*, 2016] is employed as the backbone network in our implementation. We denote the candidate bounding boxes of the human and object by b_h and b_o, respectively. We also express the box of the whole scene as b_s. After the process of shared ResNet-50 C1-C4 and RoI pooling according to respective candidate boxes, ResNet-50 C5 generates input features for targets T_{scene}, T_{human}, and T_{object}, teamed as f_s, f_h and f_o. These three targets are then assigned into three in-Graph models in pairs. While input features for in-Graph model have dimension $D = 2048$, we set reduced dimension $C = 1024$, number of nodes $N = 512$.

Three Branches

The in-GraphNet contains a human-centric branch, an object-centric branch and a spatial configuration branch, where f_h, f_o and the spatial configuration [Gao *et al.*, 2018] based on two-channel interaction pattern [Chao *et al.*, 2018] are adopted as basic feature representations, respectively. In scene-wide interactive reasoning, semantic features obtained from human-scene in-Graph and object-scene in-Graph are concatenated to f_h and f_o respectively to enrich presentations in human-centric and object-centric branches. In instance-wide interactive reasoning, semantic feature output from human-object in-Graph is concatenated into the object-centric branch only, because appearance of an object is usually constant in different interactions and provides minor effect in human-centric representation. Finally, the enhanced features from three branches are fed into fully connected layers and perform classification operations. In this way, the entire framework is implemented to be fully differentiable and end-to-end trainable using gradient-based optimization. With the formulation above, rich interactive relations between visual targets can be explicitly utilized to infer HOIs.

Training

Since HOI detection is a multi-label classification problem where more than one HOI label might be assigned to a $\langle$ human, object $\rangle$ candidate, our model is trained in a supervised fashion using the multi-label binary cross-entropy loss. All three branches are trained jointly, where the overall loss for each interaction category is the sum of three losses from three branches.

Method	Backbone Network	mAP$_{role}$(%)
Gupta et al. [Yatskar et al., 2016]	ResNet-50-FPN	31.8
InteractNet [Gkioxari et al., 2018]	ResNet-50-FPN	40.0
BAR-CNN [Kolesnikov et al., 2018]	Inception-ResNet	41.1
GPNN [Qi et al., 2018]	Deformable CNN	44.0
iCAN [Gao et al., 2018]	ResNet-50	45.3
Contextual Att [Wang et al., 2019]	ResNet-50	47.3
iHOI [Xu et al., 2019]	ResNet-50	40.4
RPNN [Zhou and Chi, 2019]	ResNet-50	47.5
TIN(R$P_d C_d$) [Li et al., 2019]	ResNet-50	47.8
our baseline	ResNet-50	44.8
In-GraphNet (ours)	ResNet-50	**48.9**

Table 1: Performance comparison with state-of-the-arts on V-COCO dataset.

Method	Backbone Network	Default		
		full	rare	non-rare
Shen et al. [Shen et al., 2018]	VGG-19	6.46	4.24	7.12
HO-RCNN [Chao et al., 2018]	CaffeNet	7.81	5.37	8.54
InteractNet [Gkioxari et al., 2018]	ResNet-50-FPN	9.94	7.16	10.77
GPNN [Qi et al., 2018]	Deformable CNN	13.11	9.34	14.23
iCAN [Gao et al., 2018]	ResNet-50	14.84	10.45	16.15
Contextual Att [Wang et al., 2019]	ResNet-50	16.24	11.16	17.75
iHOI [Xu et al., 2019]	ResNet-50	9.97	7.11	10.83
RPNN [Zhou and Chi, 2019]	ResNet-50	17.35	12.78	18.71
TIN(R$P_d C_d$) [Li et al., 2019]	ResNet-50	17.03	**13.42**	
our baseline	ResNet-50	15.41	10.71	16.81
In-GraphNet (ours)	ResNet-50	**17.72**	12.93	**19.31**

Table 2: Results on HICO-DET test set. Default: all images. Full: all 600 HOI categories. Rare: 138 HOI categories with less than 10 training instances. Non-Rare: 462 HOI categories with 10 or more training instances.

For each image, pairwise candidate boxes ($B_h * B_o$) for each HOI category are assigned binary labels based on the prediction. Similar to general HOI detection frameworks [Gkioxari et al., 2018; Gao et al., 2018], we use a fusion of scores output from each branch to predict a final score for each HOI.

$$S_{h,o} = S_h^{pre} * S_o^{pre} * (S_h + S_o) * S_s. \qquad (4)$$

Here S_h, S_o and S_s denotes scores output from binary sigmoid classifiers in human-centric branch, object-centric branch and spatial configuration branch, respectively. S_h^{pre} and S_o^{pre} are object detection scores output from object detector for candidate boxes. $S_{h,o}$ is the final score for each HOI.

4 Experiments and Evaluations

4.1 Experimental Setup

Datasets and evaluation metrics. We evaluate our model and compare it with the state-of-the-arts on two large-scale benchmarks, including V-COCO [Yatskar et al., 2016] and HICO-DET [Chao et al., 2018] datasets. V-COCO [Yatskar et al., 2016] includes 10,346 images, which is a subset of MS COCO dataset. It contains 16,199 human instances in total and provides 26 common HOI annotations. HICO-DET [Chao et al., 2018] contains about 48k images and 600 HOI categories over 80 object categories, which provides more than 150K annotated ⟨ human, object ⟩ pairs. We use role mean average precision (role mAP) [Yatskar et al., 2016] on both benchmarks.
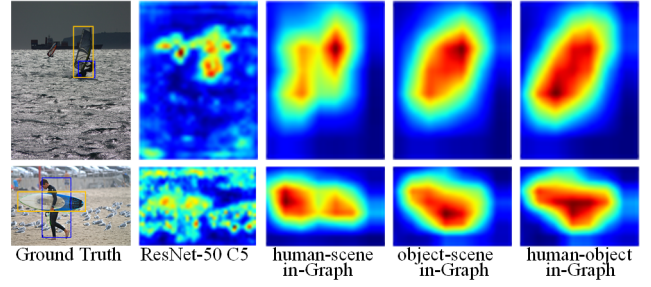
Figure 4: in-Graph visualization. The first column shows two HOI examples (i.e., ⟨ human, surf, surfboard ⟩, ⟨ human, hold, surfboard ⟩). The rest of columns visualize feature responses from ResNet-50 C5, human-scene in-Graph, object-scene in-Graph and human-object in-Graph, respectively.

number of nodes (N)	128	256	512	1024
mAP$_{role}$(%)	47.7	48.2	48.9	48.7

Table 3: Building interactive spaces with different number of nodes.

Implementation details. Following the protocol in [Gao et al., 2018], human and object bounding boxes are generated using the ResNet-50 version of Faster R-CNN [Ren et al., 2017]. Human boxes with scores higher than 0.8 and object boxes with scores higher than 0.4 are kept for detecting HOIs. We train our model with Stochastic Gradient Descent (SGD), using a learning rate of 1e-4, a weight decay of 1e-4, and a momentum of 0.9. The strategy of interactiveness knowledge training [Li et al., 2019] is adopted in our training and the model is trained for 300K and 1800K iterations on V-COCO and HICO-DET, respectively. All our experiments are conducted by tensorflow on a GPU of GeForce GTX TITAN X.

4.2 Overall Performance

We compare our method with several state-of-the-arts in this subsection. Methods being compared are classified into two categories, i.e., methods that are free from human pose (top of tables) and methods relying on additional pose estimators (middle of tables). Meanwhile, we strip all modules related to in-Graphs from our proposed framework as the baseline. Comparison results on V-COCO and HICO-DET in terms of mAP_{role} are shown in Table 1 and Table 2, respectively.

Firstly, our in-GraphNet obtains 48.9 mAP on V-COCO and 17.72 mAP on HICO-DET (Defualt full mode) and achieves absolute gains of 4.1 points and 2.3 points compared with the baseline, which are relative improvements of 9.4% and 15%. Besides, although pose-based methods usually perform better than pose-free methods, our method as a pose-free one outperforms all the others with the best performance, validating its efficacy for HOI detection task.

4.3 Ablation Studies

We adopt several ablation studies in this subsection. VCOCO serves as the primary testbeds on which we further analyze the individual effect of components in our method.
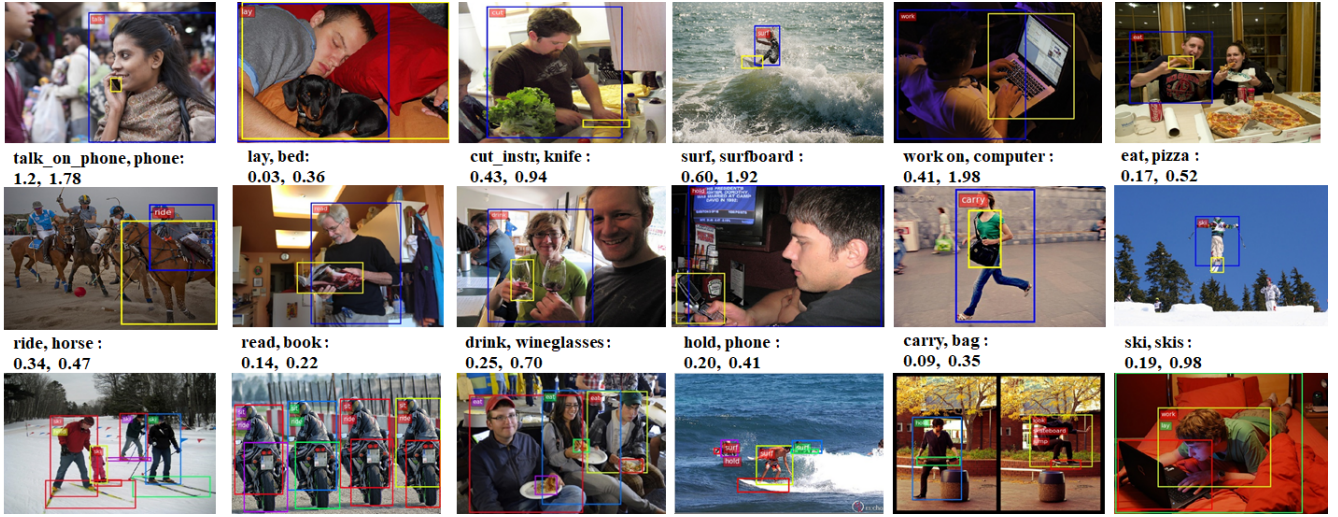
Figure 5: Visualization of HOI detections. The first and second rows show results compared with baseline on V-COCO. Each subplot displays one detected ⟨ human, verb, object ⟩ triplet for easy observation. Texts below indicate the ⟨ verb, object ⟩ tuple and two numbers in turn represent scores predicted by baseline and our approach. The last row shows multiple people take interactions with various objects concurrently detected by our method.

feature concatenation		$\sqrt{}$			
scene-wide in-Graphs			$\sqrt{}$		$\sqrt{}$
instance-wide in-Graph				$\sqrt{}$	$\sqrt{}$
mAP$_{role}$(%)	44.8	45.2	48.3	47.7	48.9

Table 4: Impact of adopting different in-Graph models.

Comprehending in-Graph. We show two examples in Figure 4 to visualize the effects of three in-Graphs been adopted. The brightness of the pixel indicates how much the feature been noticed. Intuitively, the three in-Graphs learn different interactive semantics from pairwise reasoning. Human-scene in-Graph and Object-scene in-Graph exploit interactive semantics between scene and instances. Human-object in-Graph, on the other hand, mostly focuses on the regions roughly correspond to the on-going action between human and object. In addition, to further dissect in-Graph model, empirical tests have been conducted to study the effect of number of nodes in interactive spaces. As summarized in Table 3, we get the best result when N is set as the value of 512.

Effects of adopting different in-Graph models. As shown in Table 4, while directly concatenating $[f_h, f_s]$ in human-centric branch and $[f_o, f_s, f_h]$ in object-centric branch, we can see that simply concatenating different targets without interactive reasoning barely improves the result. When we only adopt the scene-wide in-Graphs or the instance-wide in-Graph, the mAP are 48.3 and 47.7 respectively, indicating the respective effects of these two parts. Specifically, we can draw a relatively conclusion from detailed class-wise results that scene-wide in-Graphs are more adept in modelling interactions closely related to the environment, while instance-wide in-Graph performs better in depicting interactions closely related to human pose.

4.4 Quantitive Examples

For visualization, several examples of detection are given in Figure 5. We first compare our results with baseline to demonstrate our improvements. We can see from the first two rows that our method is capable of detecting various HOIs with higher scores. In addition, the third row shows that our method can adapt to complex environments to detect multiple people taking different interactions with diversified objects.

5 Conclusion

In this paper, we propose a graph-based model to address the problem of lack interactive reasoning in existing HOI detection methods. Beyond convolutions, our proposed in-Graph model efficiently reasons interactive semantics among visual targets by three procedures, i.e., a project function, a message passing process and an update function. We further construct an in-GraphNet assembling two-level in-Graph models in a multi-stream framework to parse scene-wide interactive semantics and instance-wide interactive semantics for inferring HOIs. The in-GraphNet is free from costly human pose and end-to-end trainable. Extensive experiments have been conducted to evaluate our method on two public benchmarks, including V-COCO and HICO-DET. Our method outperforms both existing human pose-free and human pose-based methods, validating its efficacy in detecting HOIs.

Acknowledgments

References

[Chao *et al.*, 2018] Yuwei Chao, Yunfan Liu, Xieyang Liu, Huayi Zeng, and Jia Deng. Learning to detect human-object interactions. *workshop on applications of computer vision*, pages 381–389, 2018.

[Fang *et al.*, 2017] Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu. Rmpe: Regional multi-person pose estimation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2334–2343, 2017.

[Fang *et al.*, 2018] Hao-Shu Fang, Jinkun Cao, Yu-Wing Tai, and Cewu Lu. Pairwise body-part attention for recognizing human-object interactions. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 51–67, 2018.

[Gao *et al.*, 2018] Chen Gao, Yuliang Zou, and Jiabin Huang. ican: Instance-centric attention network for human-object interaction detection. *british machine vision conference*, page 41, 2018.

[Gilmer *et al.*, 2017] Justin Gilmer, Samuel S Schoenholz, Patrick F Riley, Oriol Vinyals, and George E Dahl. Neural message passing for quantum chemistry. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1263–1272. JMLR. org, 2017.

[Gkioxari *et al.*, 2015] Georgia Gkioxari, Ross B Girshick, and Jitendra Malik. Contextual action recognition with r*cnn. *international conference on computer vision*, pages 1080–1088, 2015.

[Gkioxari *et al.*, 2018] Georgia Gkioxari, Ross B Girshick, Piotr Dollar, and Kaiming He. Detecting and recognizing human-object interactions. *computer vision and pattern recognition*, pages 8359–8367, 2018.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[He *et al.*, 2017] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[Idrees *et al.*, 2017] Haroon Idrees, Amir R Zamir, Yu-Gang Jiang, Alex Gorban, Ivan Laptev, Rahul Sukthankar, and Mubarak Shah. The thumos challenge on action recognition for videos "in the wild". *Computer Vision and Image Understanding*, 155:1–23, 2017.

[Kolesnikov *et al.*, 2018] Alexander Kolesnikov, Christoph H Lampert, and Vittorio Ferrari. Detecting visual relationships using box attention. *arXiv: Computer Vision and Pattern Recognition*, 2018.

[Li *et al.*, 2018] Qimai Li, Zhichao Han, and Xiao-Ming Wu. Deeper insights into graph convolutional networks for semi-supervised learning. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[Li *et al.*, 2019] Yong-Lu Li, Siyuan Zhou, Xijie Huang, Liang Xu, Ze Ma, Hao-Shu Fang, Yanfeng Wang, and Cewu Lu. Transferable interactiveness knowledge for human-object interaction detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3585–3594, 2019.

[Qi *et al.*, 2018] Siyuan Qi, Wenguan Wang, Baoxiong Jia, Jianbing Shen, and Songchun Zhu. Learning human-object interactions by graph parsing neural networks. *european conference on computer vision*, pages 407–423, 2018.

[Ren *et al.*, 2017] Shaoqing Ren, Kaiming He, Ross B Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6):1137–1149, 2017.

[Sharma *et al.*, 2015] Shikhar Sharma, Ryan Kiros, and Ruslan Salakhutdinov. Action recognition using visual attention. *international conference on learning representations*, 2015.

[Shen *et al.*, 2018] Liyue Shen, Serena Yeung, Judy Hoffman, Greg Mori, and Li Feifei. Scaling human-object interaction recognition through zero-shot learning. *workshop on applications of computer vision*, pages 1568–1576, 2018.

[Wan *et al.*, 2019] Bo Wan, Desen Zhou, Yongfei Liu, Rongjie Li, and Xuming He. Pose-aware multi-level feature network for human object interaction detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 9469–9478, 2019.

[Wang *et al.*, 2019] Tiancai Wang, Rao Muhammad Anwer, Muhammad Haris Khan, Fahad Shahbaz Khan, Yanwei Pang, Ling Shao, and Jorma Laaksonen. Deep contextual attention for human-object interaction detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5694–5702, 2019.

[Xu *et al.*, 2019] Bingjie Xu, Junnan Li, Yongkang Wong, Qi Zhao, and Mohan S Kankanhalli. Interact as you intend: Intention-driven human-object interaction detection. *IEEE Transactions on Multimedia*, 2019.

[Yang *et al.*, 2018] Maoke Yang, Kun Yu, Chi Zhang, Zhiwei Li, and Kuiyuan Yang. Denseaspp for semantic segmentation in street scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3684–3692, 2018.

[Yatskar *et al.*, 2016] Mark Yatskar, Luke Zettlemoyer, and Ali Farhadi. Situation recognition: Visual semantic role labeling for image understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5534–5542, 2016.

[Zhou and Chi, 2019] Penghao Zhou and Mingmin Chi. Relation parsing neural network for human-object interaction detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 843–851, 2019.