Cascaded Parsing of Human-Object Interaction Recognition

Tianfei Zhou, Siyuan Qi, Wenguan Wang, *Member IEEE*,
Jianbing Shen, *Senior Member IEEE*, Song-Chun Zhu, *Fellow IEEE*

Abstract—This paper addresses the task of detecting and recognizing human-object interactions (HOI) in images. Considering the intrinsic complexity and structural nature of the task, we introduce a cascaded parsing network (CP-HOI) for a multi-stage, structured HOI understanding. At each cascade stage, an instance detection module progressively refines HOI proposals and feeds them into a structured interaction reasoning module. Each of the two modules is also connected to its predecessor in the previous stage, enabling efficient cross-stage information propagation. The structured interaction reasoning module is built upon a graph parsing neural network (GPNN), which efficiently models potential HOI structures as graphs and mines rich context for comprehensive relation understanding. In particular, GPNN infers a parse graph that i) interprets meaningful HOI structures by a learnable adjacency matrix, and ii) predicts action (edge) labels. Within an end-to-end, message-passing framework, GPNN blends learning and inference, iteratively parsing HOI structures and reasoning HOI representations (*i.e.*, instance and relation features). Further beyond relation detection at a bounding-box level, we make our framework flexible to perform fine-grained pixel-wise relation segmentation; this provides a new glimpse into better relation modeling. A preliminary version of our CP-HOI model reached 1^{st} place in the ICCV2019 Person in Context Challenge, on both relation detection and segmentation. In addition, our CP-HOI shows promising results on two popular HOI recognition benchmarks, *i.e.*, V-COCO and HICO-DET.

Index Terms—Human-Object Interaction Recognition, Cascaded Parsing, Fine-Grained Relation Segmentation

1 Introduction

HUMAN-object interaction (HOI) recognition aims to identify meaningful ⟨*human, verb, object*⟩ triplets from images, such as ⟨*human, read, laptop*⟩ in Fig. 1. It plays a crucial role in many vision tasks, *e.g.*, visual question answering [3]–[5], human-centric understanding [6]–[8], image generation [9], and activity recognition [10]–[15], to name a few representative ones. Beyond the traditional visual recognition of individual instances, *e.g.*, human pose estimation, action recognition, and object detection, recognizing HOIs requires a deeper semantic understanding of image content.

A successful HOI recognition model must accurately **1)** localize and recognize each interacting entity instance (*human, object*), and **2)** predict the interaction classes (*verb*). Since both subtasks are difficult, HOI recognition is a challenging problem. In this paper, we address HOI recognition through a novel cascaded parsing framework (CP-HOI). CP-HOI is able to learn effective instance and mutual relation representations by cascaded refinement, and parse complex HOI structures by comprehensive inference. **First**, with a broader view of other computer vision and machine learning related fields, coarse-to-fine and cascade algorithms have been shown to deal well with complex problems [16]–[18]. The central idea is to leverage sequences of increasingly fine approximations to improve learning and inference. This motivates us to build our HOI recognition model in a

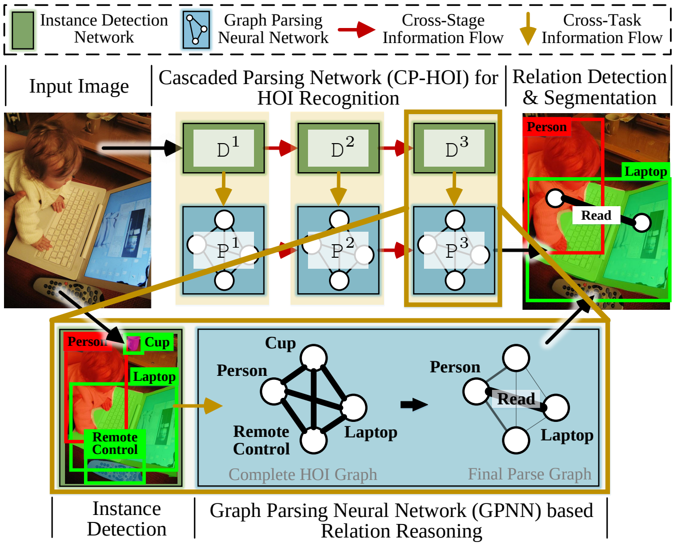

Instance Detection Network
Graph Parsing Neural Network
Cross-Stage Information Flow
Cross-Task Information Flow
Input Image
Cascaded Parsing Network (CP-HOI) for HOI Recognition
Relation Detection & Segmentation
D1
D2
D3
P1
P2
P3
Person
Laptop
Read
Person
Cup
Laptop
Remote Control
Cup
Person
Remote Control
Laptop
Complete HOI Graph
Person
Read
Laptop
Final Parse Graph
Instance Detection
Graph Parsing Neural Network (GPNN) based Relation Reasoning

Fig. 1. Our proposed CP-HOI is able to handle both object-level relation detection and pixel-wise relation segmentation. Given an input image, CP-HOI performs coarse-to-fine inference over instance detection ($\mathtt{D}^1 \sim \mathtt{D}^3$) and structured interaction recognition ($\mathtt{P}^1 \sim \mathtt{P}^3$). The interaction recognition is achieved by a graph parsing neural network (GPNN), which alternatively infers meaningful HOI structures and propagates information over the structures. The final parse graph explains the given scene with the connectivity between nodes (*e.g.*, the strong edge between the person and laptop) and the edge labels (*e.g.*, read).

- *T. Zhou and W. Wang are with ETH Zurich, Switzerland. (Email: {ztfei.debug, wenguanwang.ai}@gmail.com)*
- *S. Qi is with Google, USA. (Email: syqi@cs.ucla.edu)*
- *J. Shen is with Inception Institute of Artificial Intelligence, UAE. (Email: shenjianbingcg@gmail.com)*
- *S.-C. Zhu is with Tsinghua University and Peking University, Beijing, China.*
- *This work builds upon two earlier conference papers, appeared in ECCV2018 [1] and CVPR2020 [2], respectively.*
- *Corresponding author: Wenguan Wang*

cascaded manner, with multiple stages to more accurately identify entities and learn effective HOI representations in an annealing-style. **Second**, due to the structured nature of the task, it is desired to make use of the learned HOI representations from a comprehensive view, for better mining the rich context and understanding the trivial interactions among the individual entities. Considering this, we further propose a graph parsing neural network (GPNN), which explicitly parses meaningful HOI structures, while simultaneously reasoning the relations over the learned structures in a feedback manner. **Third**, for the two subtasks of instance detection and GPNN-based interaction recognition, CP-HOI arranges them in a successive manner within each individual stage, and carries out multi-step, cross-stage inference for each. All the above designs result in a multi-task, coarse-to-fine HOI recognition framework, which enables asymptotically improved HOI representation learning as well as effective HOI structure understanding.

In particular, as shown in Fig. 1, our CP-HOI consists of an instance detection module (🟩) and a GPNN-based interaction recognition module (🟦), both working in a cascade manner. Through the instance localization network, CP-HOI increases the selectiveness of the instance proposals step-by-step. With such progressively refined HOI candidates, more powerful HOI representations can be captured to further better support structured HOI reasoning. In addition, GPNN represents all the detected instances and potential human-object relations as a complete graph and models the HOI recognition as a parse graph inference problem. The meaningful HOI structures are interpreted as a learnable adjacency matrix, while edge types, which are identified from the parsing graph, correspond to the inferred interactions. Our approach is thus efficient, as it blends learning and inference over structures in an end-to-end manner, and makes use of cascade paradigms for multi-step refinement.

More essentially, previous HOI literature mainly addresses *relation detection*, *i.e.*, recognizing HOIs at a bounding-box level. In addition to addressing this classic setting, we take a further step towards more fine-grained HOI understanding, *i.e.*, identifying the relations between interacting entities at the pixel level (see Fig. 1). Studying such a *relation segmentation* setting not only demonstrates the efficacy and flexibility of our model, but allows us to explore more powerful HOI representations. This is because bounding-box based representations only encode coarse object information with noisy backgrounds, while pixel-wise mask based features may capture more detailed and precise cues. By extending our CP-HOI model to the relation segmentation setting, we empirically study the effectiveness of bounding-box and pixel-wise mask based relation representations as well as their hybrids. Our results suggest that the pixel-mask representation is indeed more powerful.

Our main **contributions** are summarized as follows:

- We proffer a cascaded deep structured model, CP-HOI, which addresses efficient HOI representation learning, as well as comprehensive relation reasoning simultaneously.
- The cascaded network design empowers CP-HOI with a high learning capacity, giving rise to more precise HOI representation modeling. Both the entity detection and relation reasoning are arranged in a cascaded manner and are closely coupled within each cascade stage, yielding a compact yet powerful HOI reasoning framework.
- Benefiting from the complementary strengths of neural networks and graphical models, CP-HOI is able to explicitly model individual entities and mutual relations as graphs and efficiently parse the optimal HOI structures/ineteraction labels in an end-to-end manner. It iteratively infers the structures and, in turn, broadcasts information over the learned structure, enabling a more complete and structured HOI understanding.
- By extending our CP-HOI model to pixel-level HOI understanding, which has never been touched before, we show the high flexibility and strong generalizability of our framework. In addition, to the best of our knowledge, this work provides the first effective demonstration of fine-grained HOI representation learning, which provides an insightful glimpse into the task.

The main modules in our CP-HOI model were adopted in the winning entry of the **ICCV-2019 Person in Context Challenge**[1] (PIC$_{19}$ Challenge), for both *Human-Object Interaction in the Wild (HOIW)* and *Person in Context (PIC)* tracks. HOIW addresses relation detection, while PIC focuses on relation segmentation. Besides, we evaluate the efficacy of CP-HOI on the V-COCO [19] and HICO-DET [20] datasets. Overall, CP-HOI consistently achieves promising results over the four datasets and two different settings, which reveals its remarkable performance and strong generalization.

This work builds upon two earlier conference papers [1], [2]. In [1][2], we proposed GPNN, making the first attempt to apply neural networks for both structured HOI modeling as well as iterative inference. In [2][2], we presented a cascaded HOI recognition framework (C-HOI) that, for the first time, addresses the task in an end-to-end, cascaded architecture. These papers have led to several follow-up works by ourselves [12] and other groups [21]–[26]. For the present paper, we develop a more powerful HOI recognition model, CP-HOI, which integrates the structured HOI relation reasoning and progressive HOI representation learning in a unified, cascade framework. Thus, CP-HOI inherits the advantages of GPNN and C-HOI. Based on this, we further consolidate the overall technique. In addition, the GPNN is improved with more powerful relation features as well as an edge-centric message-passing procedure. Moreover, we provide more comprehensive experiments, thoroughly examining the effectiveness of our GPNN, C-HOI and CP-HOI models.

2 Related Work

In this section, we briefly review the literature in three related fields: human-object interaction recognition (§2.1), cascade neural networks (§2.2), and graph neural networks (§2.3).

2.1 Human-Object Interaction Recognition

Reasoning human actions with objects (like "playing baseball", "playing guitar"), rather than recognizing individual actions ("playing") or object instances ("baseball", "guitar"), is essential for a more comprehensive understanding of

[1] http://picdataset.com/challenge/leaderboard/pic2019

[2] GPNN: https://github.com/SiyuanQi/gpnn
C-HOI: https://github.com/tfzhou/C-HOI

what is happening in the scene. Such a task requires *reasoning* beyond *perception*, by integrating information from human, objects, and their complex relationships. It has a rich history in computer vision. Early methods were mainly built upon structured models. Specifically, they studied the Bayesian model [27], [28], utilized contextual relationships between humans and objects [29], [30], learned structured representations with spatial interactions and context [31], exploited compositional models [32], or referred to a set of HOI exemplars [33]. Though remarkable results were achieved, these methods require carefully hand-designed pipelines and suffer the limited representability of hand-crafted features (*e.g.*, color, HOG, and SIFT).

With the recent renaissance of neural networks in computer vision and the availability of large-scale HOI datasets [19], [20], [34], deep learning based solutions are now dominant in this field. For instance, Mallya *et al*. [35] modified Fast R-CNN [36] for HOI recognition, with the assistance of visual question answering. In [37], a multi-branch architecture was explored to address human, object, and relation representation learning. Based on this idea, Gkioxari *et al*. [38] further equipped Faster R-CNN [39] with a new branch for interaction modeling. To learn more effective HOI representations, recent leading approaches widely made use of pose cues [23], [26], [40]–[42], tried to automatically discover informative human parts [43], or leveraged context information from the background [44]. Some other efforts addressed long-tail distribution and zero-shot problems with external knowledge [45]–[48] or transfer learning [49], [50]. Although promising results have been achieved by these deep HOI models, there still remain two unsolved issues. First, they lack a powerful tool to explicitly represent the structures in the HOI task. Second, all these models are built upon single-stage pipelines for learning and inference, which are limited in handling the inherent challenges in the task. In contrast, we propose a general architecture that addresses coarse-to-fine feature learning as well as structured HOI relation reasoning within a multi-stage pipeline. In addition, previous efforts in HOI recognition mainly focused on understanding human-object relations at a bounding-box level. In this work, with the release of PIC$_{19}$, we propose the first solution towards pixel-level HOI understanding and further study the effectiveness of different HOI representations (*i.e.*, fine-grained, pixel-wise masks and classic, bounding boxes).

2.2 Cascade Neural Networks

Cascade is one of the most classic and powerful algorithms in computer vision. The essence of cascade is to build more discriminative classifiers by stacking multi-stage classifiers, such that early stages discard a large number of easy negative samples so that later stages can focus on handling more difficult examples [51]. Such an idea has been explored in various forms, with a history dating back to the 1970's (as pointed out by [52]). Cascade methods based on hand-crafted features have shown wide success in various tasks, such as generic object detection [17] and face detection [53].

Recently, several efforts have been made towards endowing deep neural networks (DNN) with cascade architectures. Specifically, current popular two-stage object detectors [39], [54]–[56] can be considered as cascade models, where the background regions are removed in the first stage and the remaining object proposals are further refined in the second stage. Some researchers also explored the end-to-end learning of more than two cascaded DNNs and achieved promising results on several tasks, including generic object detection [54], [57], and instance segmentation [56]. In these methods, the stages are usually all supervised simultaneously to end-to-end learn gradually improved features and facilitate performance in a step-by-step manner. Another representative work is Hourglass [58], which is a well-known multi-stage pose estimator, with the particular advantage in modeling long-term dependency relations among human parts. In this work, we revisit the general idea of the cascade architecture in HOI recognition. We couple instance localization and relation recognition as a cascaded framework, which addresses the inherent challenges in this task by coarse-to-fine learning HOI representations and comprehensively parsing HOI structures.

2.3 Neural Networks with Graphs

DNNs are able to learn flexible data features, but lack intuitive high-level structures. Alternatively, graphical models are powerful at building structured representations, but often require significant feature engineering. Therefore, in the literature, several approaches have been proposed to combine graphical models and neural networks, in order to leverage their complementary advantages. The most intuitive approach is to build graphical models upon DNN, where the network that generates features is trained first, and its output is used to compute potential functions for the graphical predictor. Typical methods have been used in human pose estimation [59], human part parsing [60], and semantic image segmentation [61]. However, these methods lack a deep integration in the sense that the computation process of graphical models cannot be learned end-to-end.

An essential branch of DNNs [62], [63], *i.e.*, graph neural networks (GNNs), was developed to allow end-to-end learning over graphs and has gained widespread attention in recent years. One typical trend is to generalize classic convolution operations directly from Euclidean data (*e.g.*, images) to graphs, called Graph Convolutional Networks (GCNs). Due to the simple architecture, GCNs demonstrate advantages in handling massive graph data (*e.g.*, millions of nodes), but suffer from limited modeling ability for complex structures [63]. Another category of GNNs [64], [65], called Message Passing Graph Networks (MPGNs), are more related to ours, which model the graph elements (*i.e.*, node, edges) and approximation inference as learnable neural networks. MPGNs typically yield complex structures, while gaining higher flexibility and learning capacity.

GNNs have obtained wide success in many fields, including molecular biology [64], computer vision [66]–[72], and machine learning [65], demonstrating their advantages in structured modeling. In this paper, we extend previous graphical neural networks with learnable graph structures, which effectively addresses the rich and high-level relations in HOI problems. The proposed GPNN can automatically infer the meaningful HOI structure and utilize that structure to enhance information propagation and facilitate further inference. Although [67], [70] also leverage GNNs to address the relation modeling among visual elements for generating scene graph, our method is unique in: 1) inferring a parse graph for explicitly modeling HOI structures, 2) exploring a comprehensive set of relational cues; and 3) adopting a cascaded learning regime for both visual instance and relation representation modeling. It is worth mentioning that some recent efforts [49], [73] also explore edge-embedding learning of graphs, for better modeling structural interactions between human and objects. Specifically, for interaction relation modeling, [49] considers both compositional and visual phrase representations, while [73] addresses the importance of spatial configurations. Our model is flexible enough to encode these cues and goes one step further to perform multi-step reasoning for relation understanding and interacting entity localization.

3 Our Algorithm

3.1 Method Overview

3.1.1 Problem Formulation

To identify $\langle$*human, verb, object*$\rangle$ triplets in images, it is desired to accurately 1) localize and recognize individual instances (*human*, *object*), and 2) reason the interacted relations (*verb*). To achieve this, we tackle the task through a cascaded and structured reasoning framework, where the cascade process is carried out to address progressive refinement over instance localization, and the structured inference is approached to comprehensively mine inherent context and efficiently recognize complex relations.

Specifically, for HOI understanding, human and object entities are first identified from the input image I. Then we construct a *complete* HOI graph $\mathcal{G}=(\mathcal{V},\mathcal{E},\mathcal{Y})$ to include all the possible relationships between the detected human and objects as well as address the context among objects. The human and objects are represented by nodes $v\in\mathcal{V}$, which take unique values from $\{1,\cdots,|\mathcal{V}|\}$. The potential pairwise relations between human and objects are defined as edges $e\in\mathcal{E}$, which are two-tuples $e=(v,w)\in\mathcal{V}{\times}\mathcal{V}$. Each edge e has an output vector $\boldsymbol{y}_{v,w}\in[0,1]^{Y}$, *i.e.*, the scores of Y *verb* labels and $\mathcal{Y}=\{\boldsymbol{y}_{v,w}\}_{(v,w)\in\mathcal{E}}$. For relation recognition, we want to automatically infer a *parse* graph $g=(\mathcal{V}_g,\mathcal{E}_g,\mathcal{Y}_g)$, *i.e.*, a sub-graph of $\mathcal{G}$ where $\mathcal{V}_g\subseteq\mathcal{V}$ and $\mathcal{E}_g\subseteq\mathcal{E}$, by keeping the meaningful HOI structures and labeling the edges. Given the node features $\mathcal{X}_{\mathcal{V}}=\{\boldsymbol{x}_v\}_{v\in\mathcal{V}}$ and edge features $\mathcal{X}_{\mathcal{E}}=\{\boldsymbol{x}_{v,w}\}_{(v,w)\in\mathcal{E}}$, we want to infer the optimal parse graph g^* that best explains the data I according to a probability distribution p:

$$\begin{aligned}
g^*&=\underset{g}{\operatorname{argmax}}\;p(g|I)=\underset{g}{\operatorname{argmax}}\;\mathop{\mathbb{E}}_{\mathcal{X},\mathcal{G}}[p(\mathcal{X},\mathcal{G}|I)p(g|\mathcal{X},\mathcal{G},I)]\\
&\approx\underset{g}{\operatorname{argmax}}\;p(\mathcal{X}^*,\mathcal{G}^*|I)p(g|\mathcal{X}^*,\mathcal{G}^*)\\
&=\underset{g}{\operatorname{argmax}}\;p(\mathcal{V}_g,\mathcal{E}_g,\mathcal{Y}_g|\mathcal{X}^*,\mathcal{G}^*)\\
&=\underset{g}{\operatorname{argmax}}\;p(\mathcal{Y}_g|\mathcal{V}_g,\mathcal{E}_g,\mathcal{X}^*)p(\mathcal{V}_g,\mathcal{E}_g|\mathcal{X}^*,\mathcal{G}^*).
\end{aligned}\tag{1}$$

Here $\mathcal{X}^*,\mathcal{G}^*=\operatorname{argmax}_{\mathcal{X},\mathcal{G}}p(\mathcal{X},\mathcal{G}|I)$ and $\mathcal{X}=\{\mathcal{X}_{\mathcal{V}},\mathcal{X}_{\mathcal{E}}\}$. We use the complete HOI graph and best feature to approximate the expectation. $p(\mathcal{V}_g,\mathcal{E}_g|\mathcal{X}^*,\mathcal{G}^*)$ evaluates the graph structure, and $p(\mathcal{Y}_g|\mathcal{V}_g,\mathcal{E}_g,\mathcal{X}^*)$ is the labeling probability for the (object) nodes in the parse graph.

This formulation provides us a principled guideline for designing our CP-HOI model. To obtain the best feature $\mathcal{X}^*$, we propose a cascaded approach that performs coarse-to-fine object localization and pair-wise relation modeling. To approximate the computations of $\operatorname{argmax}_g p(\mathcal{V}_g,\mathcal{E}_g|\mathcal{X}^*,\mathcal{G}^*)$ and $\operatorname{argmax}_g p(\mathcal{Y}_g|\mathcal{V}_g,\mathcal{E}_g,\mathcal{X}^*)$, CP-HOI is equipped with a *graph parsing neural network* (GPNN), which jointly estimates the HOI structure $\operatorname{argmax}_g p(\mathcal{V}_g,\mathcal{E}_g|\mathcal{X}^*,\mathcal{G}^*)$ and conducts information diffusion based optimal labeling approximation $\operatorname{argmax}_g p(\mathcal{Y}_g|\mathcal{V}_g,\mathcal{E}_g,\mathcal{X}^*)$.

3.1.2 Cascaded Parsing Network Architecture

Our method carries out progressive refinement on instance detection and structured relation recognition at multiple stages (see Fig. 2 (a)). At each stage t, the multi-tasking is achieved by two networks. **1)** Obtain the best feature: an cascaded instance detection network $\mathtt{D}^t$ generates gradually improved *human* and *object* proposals (representations) and pair-wise relation features. **2)** Find the best parse graph: a GPNN-based relation recognition network $\mathtt{P}^t$ infers meaningful HOI structures and predicts action labels. Our CP-HOI model is organized as:

$$\begin{aligned}
\text{Cascaded Instance Detection (§3.2):}\;\; &\mathcal{O}^t,\mathcal{G}^t=\mathtt{D}^t(I,\mathcal{O}^{t-1}),\\
\text{Structured Interaction Reasoning (§3.3):}\;\; &g^t=\mathtt{P}^t(\mathcal{X}^t,\mathcal{G}^t,g^{t-1}).
\end{aligned}$$

Here $\mathtt{D}^t$ accepts the instance detection results $\mathcal{O}^{t-1}$ from $\mathtt{D}^{t-1}$ as arguments and outputs refined $\mathcal{O}^t$ as well as a more accurate complete graph $\mathcal{G}^t$. With this $\mathcal{G}^t$ and previously inferred parse graph g^{t-1}, $\mathtt{P}^t$ generates a new parse graph estimation g^t to better explain HOI structures and predict more accurate relation labels. Notably, the instance detection $\mathtt{D}^t$ and interaction reasoning $\mathtt{P}^t$ networks work closely, as $\mathtt{P}^t$ can benefit from the improved HOI graph $\mathcal{G}^t$, which is generated from $\mathtt{D}^t$.

Next we will describe in detail our instance detection network in §3.2 and GPNN-based interaction recognition network in §3.3. For notation clarity, for each network, we mainly focus on one single stage, as the network architectures are similar across different stages.

3.2 Instance Detection and HOI Graph Construction

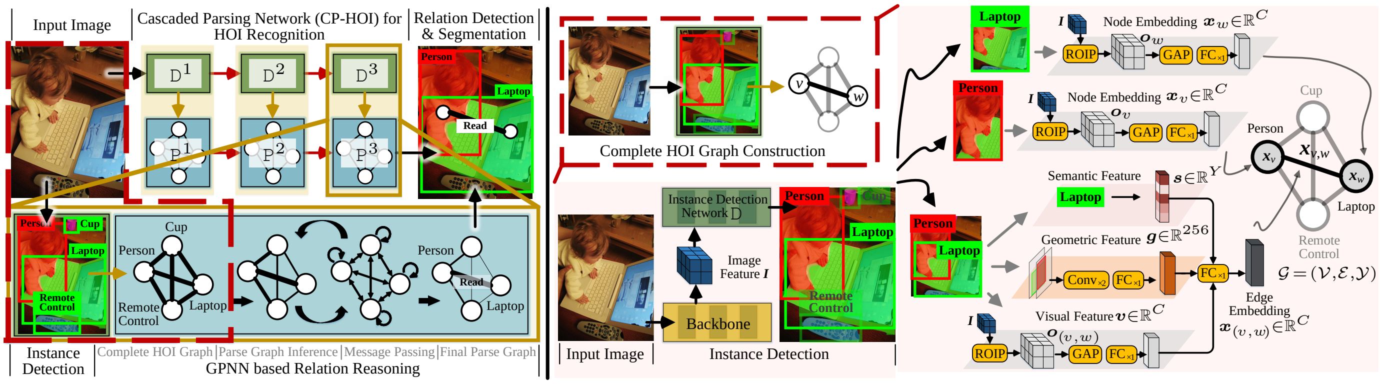
3.2.1 Cascaded Instance Localization Network

The instance detection network $\mathtt{D}$ outputs human and object candidates $\mathcal{O}$, from which a complete HOI graph $\mathcal{G}$ can be derived and fed into the interaction recognition network $\mathtt{P}$ for HOI structure parsing. Specifically, it is built on a cascade of detectors, *i.e.*, at stage t, $\mathtt{D}^t$ refines an object region $o^{t-1}\in\mathcal{O}^{t-1}$ detected from the preceding stage:

$$\boldsymbol{o}_{o^{t-1}}=\mathrm{ROIP}(\boldsymbol{I},o_{o^{t-1}}),\tag{2}$$

$$o^t=\mathtt{H}^t(\boldsymbol{o}_{o^{t-1}}),\tag{3}$$

where $\boldsymbol{I}$ is the CNN feature of the input image I, extracted by a backbone network. Further, $\boldsymbol{o}_{o^{t-1}}\in\mathbb{R}^{C\times H\times W}$ indicates the box feature derived from $\boldsymbol{I}$ and the input RoI o^{t-1}. $\mathrm{ROIP}(\cdot)$ and $\mathtt{H}^t$ represent RoIAlign [55] and a bounding box prediction head, respectively. At each cascade stage t, based on the output o^{t-1} of the previous stage t–1, the detector $\mathtt{D}$ is able to generate a more accurate object o^t with an improved representation $\boldsymbol{o}_{o^t}$, compared with $\boldsymbol{o}_{o^{t-1}}$.


Input Image
Cascaded Parsing Network (CP-HOI) for HOI Recognition
Relation Detection & Segmentation
D1
D2
D3
P1
P2
P3
Person
Laptop
Read
Cup
Remote Control
Instance Detection
Complete HOI Graph
Parse Graph Inference
Message Passing
Final Parse Graph
GPNN based Relation Reasoning
v
w
Complete HOI Graph Construction
Instance Detection Network D
Image Feature I
Backbone
Input Image
Instance Detection
Node Embedding $\boldsymbol{x}_w \in \mathbb{R}^C$
$\boldsymbol{o}_w$
Node Embedding $\boldsymbol{x}_v \in \mathbb{R}^C$
$\boldsymbol{o}_v$
ROIP
GAP
FC×1
Semantic Feature $\boldsymbol{s} \in \mathbb{R}^Y$
Geometric Feature $\boldsymbol{g} \in \mathbb{R}^{256}$
Conv×2
Visual Feature $\boldsymbol{v} \in \mathbb{R}^C$
$\boldsymbol{o}_{(v,w)}$
I
x_v
x_w
$x_{v,w}$
Edge Embedding $\boldsymbol{x}_{(v,w)} \in \mathbb{R}^C$
$\mathcal{G}=(\mathcal{V},\mathcal{E},\mathcal{Y})$

(a) Overview of our CP-HOI model during inference (b) Illustration of our instance detection module and complete HOI graph construction

Fig. 2. (a) Illustration of our cascaded parsing network (CP-HOI), which identifies a triplet of ⟨*person*, *read*, *laptop*⟩ from an input image. (b) Illustration of our instance detection module and the pipeline for constructing a complete HOI graph from detected instances (§3.2). Here we showcase how to extract initial embeddings for nodes: v (person) and w (laptop), and an edge (v, w).

3.2.2 Network Training

Similar to previous cascade object detectors [54], [56], at each cascade stage, the detector is trained with a certain interaction over union (IoU) threshold, and its output is re-sampled to train the next-stage detector with a higher IoU threshold. In this way, we gradually increase the quality of training data for deeper stages, thus boosting the selectiveness against hard negative examples. During training, at each stage t, the training loss $\mathcal{L}^t_{\mathrm{LOC}}$ for instance detection is the same as in Faster R-CNN [39].

3.2.3 Complete HOI Graph Construction

With all the detected object candidates $\mathcal{O}^t$, we next describe how contextual features $\mathcal{X}^t$ are extracted and used to instantiate nodes and edges in the complete HOI graph $\mathcal{G}^t$ (see Fig. 2 (b)). In the following paragraphs, the superscript 't' is omitted for conciseness, unless necessary.

Node Embedding Initialization. In $\mathcal{G}$, the nodes $\mathcal{V}$ are the human and object candidates $\mathcal{O}$. For each node $v \in \mathcal{V}$, its embedding $\boldsymbol{x}_v \in \mathbb{R}^C$ is initialized by:

$$\boldsymbol{x}_v = \mathrm{FC}_{\times 1}(\mathrm{GAP}(\boldsymbol{o}_v)) \in \mathbb{R}^C, \tag{4}$$

where $\mathrm{FC}_{\times 1}(\cdot)$ stands for a *fully connected layer* (FC) and $\mathrm{GAP}(\cdot)$ indicates the *global average pooling* operation. Further, $\boldsymbol{o}_v$ is the RoIAlign feature of v, computed in Eq. (2).

Edge Embedding Initialization. In $\mathcal{G}$, the edges $\mathcal{E}$ represent the pairwise relations between nodes (detected instances). For each edge $e = (v, w) \in \mathcal{E}$, three types of cues, *i.e.*, *semantic* feature $\boldsymbol{s}$, *geometric* feature $\boldsymbol{g}$ and *visual* feature $\boldsymbol{v}$, are utilized to obtain an efficient relation (edge) feature $\boldsymbol{x}_{v,w} \in \mathbb{R}^C$.

- *Semantic feature*. This captures our prior knowledge of *object affordances* [74] (*e.g.*, a phone affords calling). We build $\boldsymbol{s} \in \mathbb{R}^Y$ as the frequency of label co-occurrence between object and action categories [75], where Y denotes the number of pre-defined actions in an HOI dataset. Note that, if v and w are both objects, we directly set $\boldsymbol{s} = \mathbf{0}$.
- *Geometric feature*. This characterizes the spatial relationships between human and objects, which are informative for human-object interactions. For example, consider the *sit on* verb; from this we can deduce that the subject is above the object. Similar to [37], [43], we first adopt a two-channel mask representation strategy, obtaining a $(2, 64, 64)$-d feature tensor for the two entities. Then, two conv+pooling operations followed by a FC layer are applied on the tensor to obtain $\boldsymbol{g} \in \mathbb{R}^{256}$.
- *Visual feature*. The visual feature $\boldsymbol{v}$ is computed from the union region of v and w:

$$\boldsymbol{v} = \mathrm{FC}_{\times 1}\big(\mathrm{GAP}(\mathrm{ROIP}(\boldsymbol{I}, (v, w)))\big) \in \mathbb{R}^C. \tag{5}$$

Here $\mathrm{ROIP}(\boldsymbol{I}, (v, w))$ indicates the RoIAlign feature of the union region (v, w).

Finally, the edge embedding $\boldsymbol{x}_{v,w}$ is initialized as:

$$\boldsymbol{x}_{v,w} = \mathrm{FC}_{\times 1}([\boldsymbol{s}, \boldsymbol{g}, \boldsymbol{v}]) \in \mathbb{R}^C, \tag{6}$$

where '$[\cdot]$' indicates the concatenation operation.

So far, we have generated the human and object candidates $\mathcal{O}^t$, as well as initialized a complete HOI graph $\mathcal{G}^t$, which contains all the instances and possible interaction relations. Next, we detail our GPNN-based interaction reasoning module which infers an optimal parse graph $g^t \in \mathcal{G}^t$ to best explain the HOI structures in the given image I.

3.3 Structured Interaction Reasoning

3.3.1 Graph Parsing Neural Network

Our instance detection module D (§3.2) works in a cascade manner, and is used to approximate the complete HOI graph $\mathcal{G}^*$ and best instance and relation features $\mathcal{X}^*$ (Eq. (1)). To infer the parse graph g from the complete HOI graph $\mathcal{G}$, our structured interaction reasoning module P is designed to infer meaningful HOI structures $\mathrm{argmax}_g\, p(\mathcal{V}_g, \mathcal{E}_g | \mathcal{X}^*, \mathcal{G}^*)$ as well as predict the action labels $\mathrm{argmax}_g\, p(\mathcal{Y}_g | \mathcal{V}_g, \mathcal{E}_g, \mathcal{X}^*)$ by propagating the information over inferred HOI structures. As illustrated in Fig. 3, we introduce four types of functions as individual modules in the forward pass of a GPNN: *link functions*, *message functions*, *update functions*, and *readout functions*. The link functions estimate the graph structure, giving an approximation of $p(\mathcal{V}_g, \mathcal{E}_g | \mathcal{X}^*, \mathcal{G}^*)$. The message, update and readout functions together resemble the belief propagation process and approximate $\mathrm{argmax}_{\mathcal{Y}_g}\, p(\mathcal{Y}_g | \mathcal{V}_g, \mathcal{E}_g, \mathcal{X}^*)$.

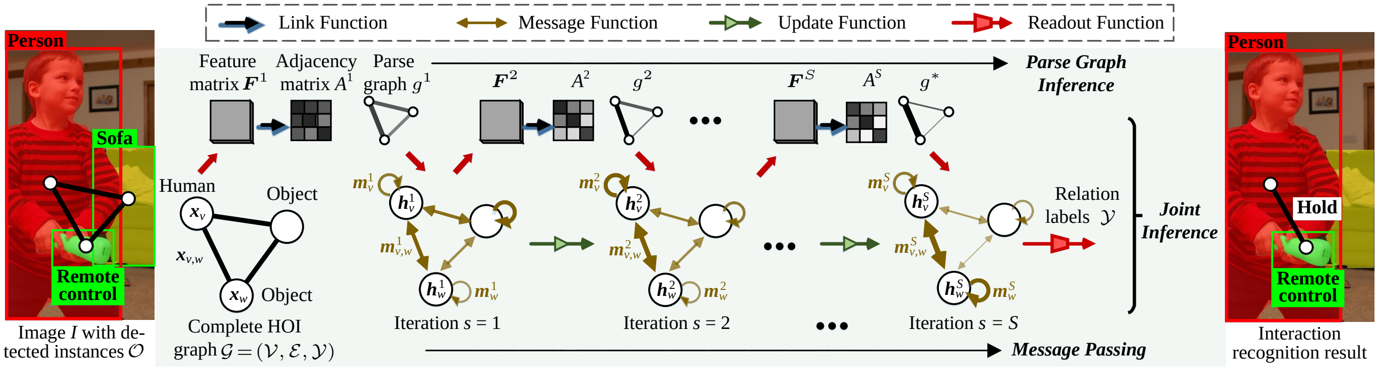


Fig. 3. Illustration of our GPNN-based structured relation reasoning (§3.3). GPNN takes node and edge features as inputs, and outputs a parse graph in a message passing fashion. The structure of the parse graph is given by a soft adjacency matrix, computed by the *link function*. The darker the color in the adjacency matrix, the stronger the connectivity is. Then *message functions* compute incoming messages for each node/edge as a weighted sum of the messages from other edges/nodes. Thicker edges indicate larger information flows. The *update functions* update the internal states of each node/edge. This process is repeated for several steps, iteratively and jointly learning the computation of HOI graph structures and message passing. Finally, for each node, the *readout functions* output HOI action labels from the hidden edge states.

In particular, the link function (➡) is used to infer the connectivities (affordance) between nodes. A soft adjacency matrix (▦) is thus constructed and used as weights for messages passing over $\mathcal{G}$. The incoming messages for nodes/edges are collected by the message function (⟷), and then fed into the update function (⇾) to update the node/edge states. Finally, the readout function (➤) computes the target outputs for each edge. These four types of functions are defined as follows:

Link Function. We first infer an adjacency matrix that represents connectivities (*i.e.*, the meaningful HOI structure) between nodes by a link function $\mathtt{L}(\cdot)$. $\mathtt{L}(\cdot)$ takes the node features $\mathcal{X}_{\mathcal{V}}$ and edge features $\mathcal{X}_{\mathcal{E}}$ as inputs, and outputs the adjacency matrix $A \in [0,1]^{|\mathcal{V}|\times|\mathcal{V}|}$:

$$A_{v,w} = \mathtt{L}(\boldsymbol{x}_v, \boldsymbol{x}_w, \boldsymbol{x}_{v,w}) \in [0,1], \tag{7}$$

where $A_{v,w}$ denotes the (v,w)-th entry of the matrix A, *i.e.*, an affordance score between nodes (instances) v and w. In this way, the structure of a parse graph g can be approximated by A. Then, we start to propagate messages over the parse graph and obtain action predictions from final edge outputs, where the soft adjacency matrix A controls the information to be passed between nodes and edges.

Message Function. During belief propagation, the hidden states of the nodes and edges are iteratively updated by communicating with other connected edges and nodes, respectively. Specifically, for each node (edge), the message functions collect information from other edges (nodes):

$$\begin{aligned} \boldsymbol{m}_v^s &= \textstyle\sum_w A_{v,w}\,\mathtt{M}_{\mathcal{V}}(\boldsymbol{h}_w^{s-1}) \in \mathbb{R}^C, \\ \boldsymbol{m}_{v,w}^s &= A_{v,w}\,\mathtt{M}_{\mathcal{E}}(\boldsymbol{h}_v^{s-1}, \boldsymbol{h}_w^{s-1}) \in \mathbb{R}^C, \end{aligned} \tag{8}$$

where $\mathtt{M}_{\mathcal{V}}(\cdot)$ and $\mathtt{M}_{\mathcal{E}}(\cdot)$ are node- and edge-specific message functions, respectively. Further, $\boldsymbol{h}_v^s$ and $\boldsymbol{h}_{v,w}^s$ are hidden states of the node v and edge $e=(v,w)$ at the s-th iteration, respectively; while $\boldsymbol{h}_v^0$ and $\boldsymbol{h}_{v,w}^0$ are set as the initial node and edge embeddings, *i.e.*, $\boldsymbol{x}_v$ and $\boldsymbol{x}_{v,w}$, respectively. Finally, $\boldsymbol{m}_v^s$ and $\boldsymbol{m}_{v,w}^s$ are messages for v and (v,w) at the s-th iteration, respectively. The connectivity A controls the information flow between nodes and edges. Eq. (8) allows information flow from edge to node and vice-versa, via multiple rounds of message passing.

Update Function. Next, the update functions update the hidden states of nodes and edges by absorbing the messages collected in Eq. (8):

$$\begin{aligned} \boldsymbol{h}_v^s &= \mathtt{U}_{\mathcal{V}}(\boldsymbol{h}_v^{s-1}, \boldsymbol{m}_v^s) \in \mathbb{R}^C, \\ \boldsymbol{h}_{v,w}^s &= \mathtt{U}_{\mathcal{E}}(\boldsymbol{h}_{v,w}^{s-1}, \boldsymbol{m}_{v,w}^s) \in \mathbb{R}^C, \end{aligned} \tag{9}$$

where $\mathtt{U}_{\mathcal{V}}(\cdot)$ and $\mathtt{U}_{\mathcal{E}}(\cdot)$ are node- and edge-specific update functions, respectively. In this way, the message-passing algorithm runs for S steps for efficient reasoning, by iteratively collecting messages (Eq. (8)) and updating node and edge embeddings (Eq. (9)).

Readout Function. After S message-passing steps, we iteratively propagate information over the HOI graph and mine richer context, thus generating more discriminative node and edge embeddings (*i.e.*, $\{\boldsymbol{h}_v^S\}_{v\in\mathcal{V}}$ and $\{\boldsymbol{h}_{v,w}^S\}_{(v,w)\in\mathcal{E}}$). For notation clarity, we rename $\boldsymbol{h}_v^S$ and $\boldsymbol{h}_{v,w}^S$ as $\boldsymbol{z}_v$ and $\boldsymbol{z}_{v,w}$, respectively. Finally, for each edge that connects a pair of human and object nodes, its final hidden state is fed into the readout function $\mathtt{R}(\cdot)$ to output an action label:

$$\boldsymbol{y}_{v,w} = \mathtt{R}(\boldsymbol{z}_{v,w}) \in [0,1]^Y. \tag{10}$$

Here, $\mathtt{R}(\cdot)$ computes the scores $\boldsymbol{y}_{v,w}$ of Y action labels for edge $e=(v,w)$ by activating its final hidden state $\boldsymbol{z}_{v,w}$. Note that here follows a multi-label classification setting.

Joint Parsing. Based on the above formulations, the messages are passed along the graph and weighted by the learned adjacency matrix A. We further extend this process into a joint learning framework that iteratively infers the graph structure and propagates the information to infer edge (interaction) labels. In particular, instead of only learning A at the beginning, we iteratively infer it with the updated node and edge features at each step s:

$$A_{v,w}^s = \mathtt{L}(\boldsymbol{h}_v^{s-1}, \boldsymbol{h}_w^{s-1}, \boldsymbol{h}_{v,w}^{s-1}). \tag{11}$$

Then the messages in Eq. (8) are redefined as:

$$\begin{aligned} \boldsymbol{m}_v^s &= \textstyle\sum_w A_{v,w}^s\,\mathtt{M}_{\mathcal{V}}(\boldsymbol{h}_w^{s-1}) \in \mathbb{R}^C, \\ \boldsymbol{m}_{v,w}^s &= A_{v,w}^s\,\mathtt{M}_{\mathcal{E}}(\boldsymbol{h}_v^{s-1}, \boldsymbol{h}_w^{s-1}) \in \mathbb{R}^C, \end{aligned} \tag{12}$$

In this way, both the graph structure and node/edge states can be jointly and iteratively learned in a unified framework.

In practice, we find such a strategy brings better performance (detailed in § 4.5).

In addition, in our cascade framework, the initial node and edge embeddings (*i.e.*, $\{\boldsymbol{x}_v^t\}_{v\in\mathcal{V}}$ and $\{\boldsymbol{x}_{v,w}^t\}_{(v,w)\in\mathcal{E}}$) at t-th cascade stage are updated by the final ones (*i.e.*, $\{\boldsymbol{z}_v^{t-1}\}_{v\in\mathcal{V}}$ and $\{\boldsymbol{z}_{v,w}^{t-1}\}_{(v,w)\in\mathcal{E}}$) in prior stage:

$$\begin{aligned}\boldsymbol{x}_v^t &\leftarrow \mathrm{FC}_{\times 2}(\boldsymbol{x}_v^t + \boldsymbol{z}_v^{t-1}) \in \mathbb{R}^C,\\ \boldsymbol{x}_{v,w}^t &\leftarrow \mathrm{FC}_{\times 2}(\boldsymbol{x}_{v,w}^t + \boldsymbol{z}_{v,w}^{t-1}) \in \mathbb{R}^C.\end{aligned} \tag{13}$$

In this way, our GPNN-based structured reasoning is also organized in a cascaded manner. At each cascade stage, benefiting from the gradually improved instance and relation features from the object detection network as well as the structural information generated by the previous GPNN module, our model performs iterative and structured reasoning for better explaining the HOI relations.

At t-th cascade stage, object candidates with low interactiveness (*i.e.*, $A_{v,w} < 0.1$) will be filtered out, before being fed to next-stage instance detection network $\mathtt{D}^{t+1}$. This helps improve the training and inference efficiency. More importantly, this lets our model focus more on those "relevant" objects and erases the difficulty caused by class imbalance, as the samples for 'non-interaction' classes are much more than the ones of any other interaction classes.

3.3.2 Network Training

Our model is built upon a message parsing based GNN, which is fully end-to-end trainable as all the functions in the iterative algorithm are differentiable neural networks. Moreover, due to its recursive nature, the model is able to process variable numbers of nodes during training and inference. When training our graph parsing neural network, two supervision signals are utilized to guide the learning of the HOI structure A, as well as edge (action) labeling $\mathcal{Y}$.

Specifically, given the complete HOI graph $\mathcal{G}$, if an edge (v, w) is annotated with an action label, we set the groundtruth value of $A_{v,w}$ as 1, to encourage the information flow among the interacting instances (*i.e.*, v and w) and corresponding edge (v, w). Otherwise, the groundtruth value of $A_{v,w}$ is set as 0, to remove the noisy information from irrelevant instances. In this way, given $\mathcal{G}$ of a training image, all the edges $\mathcal{E}$ can be further divided into two subsets: $\mathcal{E}=\hat{\mathcal{E}}\cup\check{\mathcal{E}}$, where $\hat{\mathcal{E}}$ and $\check{\mathcal{E}}$ indicate the sets of annotated and un-annotated human-object relations, respectively.

Then, at the t-th cascade stage, the training loss for learning the HOI structure A is designed as:

$$\mathcal{L}^t_{\text{structure}} = \mathcal{L}^t_{L_1} + \mathcal{L}^t_{\text{rank}}. \tag{14}$$

Here, $\mathcal{L}^t_{L_1}$ refers to the L_1 loss between A and its groundtruth. Similarly, $\mathcal{L}^t_{\text{rank}}$ is a pairwise ranking hinge loss, which is used to encourage the annotated edges $\hat{\mathcal{E}}$ to gain higher association or affordance scores than the un-annotated ones $\check{\mathcal{E}}$. This is built upon the insight that, although some human-object pairs are miss-annotated in HOI datasets, the annotated ones tend to be more relevant than those without any relation labelling. Thus, we can learn the link function $\mathtt{L}(\cdot)$ as a ranking function, which fulfills the following constraint:

$$\forall(\hat{v}, \hat{w}) \succ (\check{v}, \check{w}) : A_{\hat{v},\hat{w}} > A_{\check{v},\check{w}}, \tag{15}$$

where $(\hat{v}, \hat{w}) \in \hat{\mathcal{E}}$ and $(\check{v}, \check{w}) \in \check{\mathcal{E}}$. Note that $(\hat{v}, \hat{w}) \succ (\check{v}, \check{w})$ means $(\hat{v}, \hat{w})$ has a higher ranking than $(\check{v}, \check{w})$. For each edge (v, w), $\mathtt{L}(\cdot)$ gives the ranking score of $A_{v,w} \in [0, 1]$ by Eq. (11). Then, the ranking loss $\mathcal{L}^t_{\text{rank}}$ in Eq. (14) is achieved by:

$$\mathcal{L}_{\text{rank}} = \sum\nolimits_{(\hat{v},\hat{w})\in\hat{\mathcal{E}}} \sum\nolimits_{(\check{v},\check{w})\in\check{\mathcal{E}}} \max(0, A_{\hat{v},\hat{w}} - A_{\check{v},\check{w}} + \epsilon), \tag{16}$$

where the margin ϵ is empirically set as 0.2.

For interaction recognition, the binary cross-entropy loss $\mathcal{L}^t_{\text{CE}}$ is used to evaluate the discrepancy between the output scores from edges and corresponding groundtruth targets. For our GPNN, the overall loss at step t is computed as:

$$\mathcal{L}^t_{\text{GPNN}} = \mathcal{L}^t_{\text{structure}} + \mathcal{L}^t_{\text{CE}}. \tag{17}$$

It is worth mentioning that the affinity matrix A of $\mathcal{G}$, can be learned in an implicit manner, *i.e.*, without using the loss $\mathcal{L}^t_{\text{structure}}$ in Eq. (14). In such a case, A can be viewed as a gating or neural attention mechanism, which allows the model learn to determine how nodes and edges in the graph communicate with each other by itself, sharing a similar spirit to [65]. Experiments about different HOI structure learning strategies can be found in §4.5.

3.4 Human-Object Interaction Segmentation

So far, we have strictly followed the classic *relation detection* setting in HOI recognition [22], [23], [38], [41], *i.e.*, identify the interaction entities by bounding boxes. Now we focus on how to adapt our cascade framework to *relation segmentation*, which addresses more fine-grained HOI understanding by representing each entity at the pixel level.

Network Architecture. Inspired by [54], at each cascade stage t, an instance segmentation head $\mathtt{S}^t$ is added into the instance detection network $\mathtt{D}^t$, and the whole workflow (Eqs. (2, 3)) is changed to:

$$\begin{aligned}\text{Instance Detection: } &\boldsymbol{o}_{o^{t-1}}{=}\mathrm{ROIP}(\boldsymbol{I}, o^{t-1}), o^t{=}\mathtt{H}^t(\boldsymbol{o}_{o^{t-1}}),\\ \text{Instance Segmentation: } &\boldsymbol{o}_{o^t}{=}\mathrm{ROIP}(\boldsymbol{I}, o^t),\ \bar{o}^t{=}\mathtt{S}^t(\boldsymbol{o}_{o^t}, \boldsymbol{a}^{t-1}_{o^t}),\end{aligned} \tag{18}$$

where $\bar{o}^t \in \bar{\mathcal{O}}^t$ indicates a generated object instance mask. Further, $\boldsymbol{a}^{t-1}_{o^t}$ denotes an intermediate mask representation of o^t at stage $t-1$. It is computed recursively by forwarding the ROI o^t with all the proceeding segmentation heads:

$$\boldsymbol{a}^{t-1}_{o^t} = \mathtt{T}^{t-1}(\boldsymbol{o}_{o^t}, \boldsymbol{a}^{t-2}_{o^t}), \tag{19}$$

where $\mathtt{T}^{t-1}$ is a feature transformation module of the segmentation head $\mathtt{S}^{t-1}$. It is used to integrate bounding box and mask based features, detailed in §3.5. In this manner, direct mask information propagation is introduced between segmentation heads at different stages, leading to better segmentation results with a cascade architecture.

Then, the complete HOI graph $\mathcal{G}$ is built upon the fine-grained object instances $\bar{\mathcal{O}}$. Specifically, the human and object embeddings are initialized with finer features by applying pixel-wise ROI with instance masks. For the initial edge embedding, the geometric feature $\boldsymbol{g}$ is computed based on pixel-level masks instead of bounding boxes, and the visual feature $\boldsymbol{v}$ is computed from the masked union regions of two nodes. Then, with the mask-level $\mathcal{G}$, the GPNN (§3.3) is further appended for structured relation reasoning.

Network Training. At each cascade stage, a binary cross-entropy loss $\mathcal{L}^t_{\text{SEG}}$ is used to supervise the learning of $\mathtt{S}^t$. The training of the GPNN-based structured relation reasoning is still achieved by minimizing the loss in Eq. (17).

3.5 Detailed Network Architecture

Since both cascaded object detection and structured reasoning networks mentioned above are differentiable, our CS-HOI can be trained in an end-to-end manner. In addition, our model provides a powerful yet general framework, which allows the use of different backbone networks. For fair comparison, we use ResNet-50 [76] as the backbone, following [23], [26]. We also provide more experiments regarding the performance with different backbones.

3.5.1 Instance Detection Network

The instance localization network is a multi-stage detector, which consists of a sequence of detectors trained with gradually increased IoU thresholds. During our experiments, we use three cascade detection stages. All three detection heads have the same architecture used in Faster R-CNN [39]. Each head resamples all the regressed outputs from the previous stage and improves the detection progressively.

For interaction segmentation, the instance localization network at each cascade stage t is trained by multi-tasking, *i.e.*, jointly train segmentation and detection heads. The feature transformation layer $\mathtt{T}^{t-1}$ in Eq. (19) is implemented by a small neural network, consisting of a 1×1 convolutional layer, element-wise summation and four 3×3 convolutional layers. First, $\boldsymbol{a}_{o^t}^{t-2}$ is aligned with the ROIAlign feature $\boldsymbol{o}^t$ by the 1×1 convolutional layer, and then added to $\boldsymbol{o}^t$ through element-wise summation. Finally, the fusion feature is transformed by the four consecutive convolutional layers. More implementation details of the segmentation head can be found in [54].

3.5.2 Structured Relation Reasoning Network

Our GPNN-based relation reasoning network is also organized in a cascade manner, *i.e.*, three GPNNs are constructed over the corresponding object detectors. Each GPNN leverages the gradually improved detection results, as well as the parsing results from the previous GPNN, to infer a more accurate HOI parse graph. Within each cascade stage, four neural functions are utilized to resemble a message passing process, which iteratively performs HOI structure reasoning as well as action (edge) labelling. For each GPNN, we set the iteration steps as 3 (*i.e.*, $S=3$).

Link Function. In a message passing step s, we first concatenate all the node features (*i.e.*, $\{\boldsymbol{h}^s_v\in\mathbb{R}^C\}_{v\in\mathcal{V}}$) and all the edge features ($\{\boldsymbol{h}^s_{v,w}\in\mathbb{R}^C\}_{(v,w)\in\mathcal{E}}$) to form a feature matrix $\boldsymbol{F}^s\in\mathbb{R}^{|\mathcal{V}|\times|\mathcal{V}|\times3C}$ (■ in Fig. 3). The link function $\mathtt{L}(\cdot)$ is implemented as a small neural network with several convolutional layers (with 1×1 kernels) and a *sigmoid* activation. Then the adjacency matrix $A^{(s)}\in[0,1]^{|\mathcal{V}|\times|\mathcal{V}|}$ is computed as:

$$A^s=\sigma(\boldsymbol{W}_{\text{L}}*\boldsymbol{F}^s),\tag{20}$$

where $\boldsymbol{W}_{\text{L}}$ is the learnable parameters of the convolutional layers and $*$ denotes the convolution operation. The *sigmoid* function σ normalizes the values in A^s to [0, 1]. The essential effect of multiple convolutional layers with 1×1 kernels is similar to fully connected layers applied to each individual edge, except that the filter weights are shared by all the edges. Such an operation brings high computation efficiency.

Message Function. In our implementation, the node-specific and edge-specific update functions (*i.e.*, $\mathtt{M}_{\mathcal{V}}(\cdot)$ and $\mathtt{M}_{\mathcal{E}}(\cdot)$) are FC layers, respectively. For $\mathtt{M}_{\mathcal{E}}(\cdot)$, the two input node embeddings are first combined by the concatenation operation.

Update Function. Recurrent neural networks are natural choices for simulating the iterative update process, as used in previous works [64]. Here we apply the Gated Recurrent Unit (GRU) [77] to implement the update functions $\mathtt{U}_{\mathcal{V}}(\cdot)$ and $\mathtt{U}_{\mathcal{E}}(\cdot)$ (Eq. (9)), because of its recurrent nature and smaller amount of parameters [78].

Readout Function. The readout function $\mathtt{R}(\cdot)$ in Eq. (10) consists of two fully connected layers followed by a *sigmoid* activation function. For each edge, it computes the scores of the corresponding HOI for Y action labels.

In this way, GPNN is implemented to be fully differentiable and end-to-end trainable. The error computed by the training loss can propagate back according to the chain rule.

3.5.3 Overall Training Loss

Since all the modules mentioned above are differentiable, CP-HOI can be trained in an end-to-end manner. In the *relation detection* setting, the entire loss is computed as:

$$\begin{aligned}\mathcal{L}&=\sum\nolimits_{t=1}^{T}\mathcal{L}^t_{\text{LOC}}+\mathcal{L}^t_{\text{GPNN}}\\&=\sum\nolimits_{t=1}^{T}\alpha^t\mathcal{L}^t_{\text{LOC}}+(\beta^t\mathcal{L}^t_{\text{CE}}+\gamma\mathcal{L}^t_{L_1}+\delta\mathcal{L}^t_{\text{rank}}).\end{aligned}\tag{21}$$

Here, $\mathcal{L}^t_{\text{LOC}}$ (§3.2.2) and $\mathcal{L}^t_{\text{GPNN}}$ (Eq.(17)) are the losses for the object localization and GPNN-based relation reasoning network, defined over the t-th cascade stage, respectively. The coefficients αs, βs, γ and δ are used to balance the contributions of different cascade stages and tasks. There are three cascade stages used in our method ($T=3$), and we set $\alpha=[1,0.5,0.25]$, $\beta=[1,0.5,0.25]$, $\gamma=1$, and $\delta=0.2$. In the *relation segmentation* setting, the instance segmentation head $\mathtt{S}^t$ is injected into the network (§3.4). The corresponding instance segmentation loss $\mathcal{L}^t_{\text{SEG}}$ is further added into Eq. (21), with coefficients $[1,0.5,0.25]$.

4 Experiments

Comprehensive experiments are conducted on four datasets, *i.e.*, HOIW, PIC, V-COCO [19], and HICO-DET [20]. The first two are from the PIC_{19} Challenge, and the last two are gold standard benchmarks. For the two challenge benchmark datasets, HOIW and PIC, we report the performance of our winner entry C-HOI [2], one of the preliminary versions of our CP-HOI, as the test sets are private and evaluation server has been closed. For the rest two public datasets, V-COCO and HICO-DET, we report the performance of our GPNN [1] and C-HOI [2] models, as well as their advanced version, CP-HOI, presented in this work.

4.1 Implementation Details

Unless otherwise noted, we adopt the following training settings for all the experiments. We use ResNet-50 [76] as the backbone. The training includes two phases: 1) training the instance localization network with $\sum_{t=1}^{T}\mathcal{L}^t_{\text{LOC}}$; and then 2) jointly training the instance localization and interaction recognition networks with $\sum_{t=1}^{T}\mathcal{L}^t_{\text{LOC}}+\mathcal{L}^t_{\text{GPNN}}$ (Eq. (21)).

In the first phase, the network is initialized using the weights pre-trained on COCO [79]. The three stages are trained using gradually increased IoU thresholds

TABLE 1
Relation detection results on HOIW `test` and `val` sets in PIC$_{19}$ Challenge (§4.2).

Challenge	Team	mAP$_{rel}$
PIC$_{19}$ Challenge (*HOIW* `test`)	C-HOI [2] (**Ours**)	**66.04**
	GMVM	60.26
	FINet	56.93
	F2INet	49.13
	TIN [23]	48.64

TABLE 2
Relation segmentation results on PIC `test` and `val` sets in PIC$_{19}$ Challenge. Higher values are better. Please see §4.2 for details.

Challenge	Team	R@100 mIoU: 0.25	R@100 mIoU: 0.50	R@100 mIoU: 0.75	Mean
PIC$_{19}$ Challenge (*PIC* `test`)	C-HOI [2] (**Ours**)	**60.17**	**55.11**	**42.29**	**52.52**
	HTC+iCAN	56.21	52.32	37.49	48.67
	RelNet	53.17	49.26	32.44	44.96
	XNet	38.42	33.15	17.29	29.62

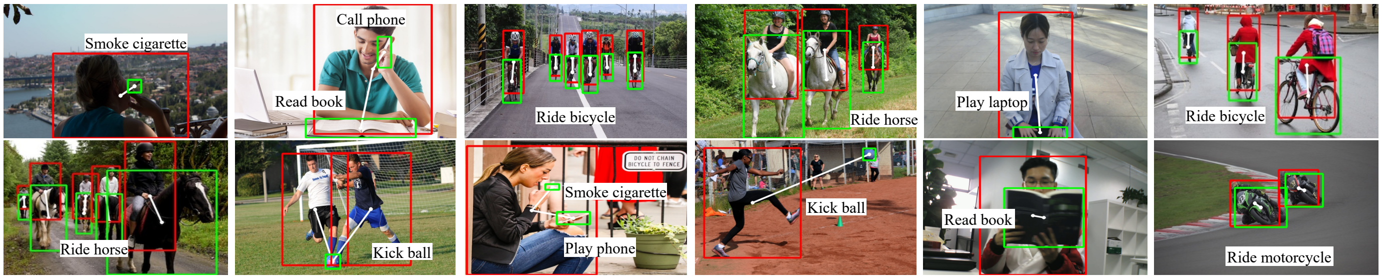


Fig. 4. Visual results for relation detection, on HOIW `test` set in PIC$_{19}$ Challenge.

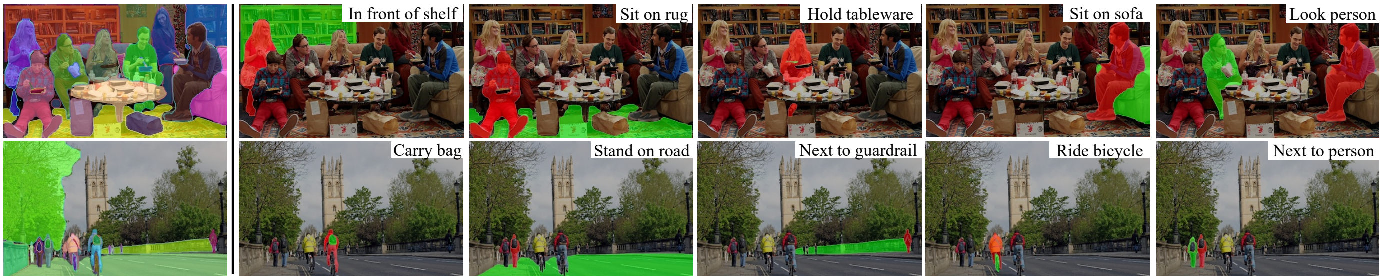


Fig. 5. Visual results for relation segmentation, on PIC `test` set in PIC$_{19}$ Challenge. First column: Instance segmentation results. Last five columns: Top ranked ⟨*human*, *verb*, *object*⟩ triplets. For each triplet, the *human* and *object* are shown in red and green.

$\{0.5, 0.6, 0.7\}$ [54], [56]. We train the network for 12 epochs with a batch size of 16 and an initial learning rate of 0.02, which is reduced by 10 at epoch 8 and 11. In the second phase, the node and edge embeddings are initialized by Eq. (4) and Eq. (6), respectively. The dimension of each node/edge feature is set as $C = 1024$. The second phase is trained with an initial learning rate of 1e-3, which is decayed by 10 at epoch 16 and 19. We train the network for 20 epoches with a batch size of 128 using four GPUs. During the two phases, training images are resized to a maximum scale of 1333×800, without changing the aspect ratio, and horizontal flipping is applied for data augmentation.

Cascade and Structured Inference. During testing, given an input image, we perform cascade and structured inference for the object detection and relation prediction simultaneously. Post-processing is conducted to obtain more accurate results: 1) objects with confidence scores smaller than 0.3 are discarded; 2) HOI scores of multiple stages are averaged, bearing similar mechanisms with model ensemble. For a test image with resolution 1333×800, the runtime is about 230ms.

Reproducibility: Our model is implemented on PyTorch and trained on eight NVIDIA Tesla V100 GPUs with 32GB memory per card. Testing is conducted on a single NVIDIA TITAN Xp GPU with 12 GB memory.

4.2 Results on PIC$_{19}$ Challenge

Dataset: The PIC$_{19}$ Challenge includes two tracks, *i.e.*, HOIW and PIC tracks, each with a standalone dataset:

- **HOIW** [34] is for human-object relation detection. It has 29,842 training and 8,794 testing images, with bounding box annotations for 11 object and 10 action categories. Since it does not provide `train/val` splits, in our ablation study, we randomly choose 9,999 images for `val` and the other 19,843 for `train`; for the challenge result, we use `train+val` for training.
- **PIC** is for human-object relation segmentation. It has 17,606 images (12,654 for `train`, 1,977 for `val` and 2,975 for `test`) with pixel-level annotations for 143 objects. It covers 30 relationships, including 6 geometric (*e.g.*, *next-to*) and 24 non-geometric (*e.g.*, *look*, *talk*).

Evaluation Metrics: Standard evaluation metrics from the challenges are adopted. For HOIW, the performance is evaluated by mAP$_{rel}$. A detected triplet ⟨*human*, *verb*, *object*⟩ is considered to be a true positive if the predicted *verb* is correct and both the *human* and *object* boxes have IoUs of at least 0.5 with the corresponding ground-truths. For PIC, we use Recall@100 (R@100), which is averaged over two relationship categories (*i.e.*, *geometric* and *non-geometric*) and

TABLE 3
Performance comparison on V-COCO `test` [19] in terms of mAP$_{role}$ (§4.3). †: an extra pose estimator is used. Higher values are better.

Methods	Publication	Backbone	mAP$_{role}$
Gupta *et. al.* [19]	Arxiv15	ResNet-50-FPN	31.8
Interact [38]	CVPR18	ResNet-50-FPN	40.0
iCAN [43]	BMVC18	ResNet-50	45.3
Xu *et. al.* [22]	CVPR19	ResNet-50-FPN	45.9
Wang *et. al.* [44]	ICCV19	ResNet-50	47.3
RPNN [26]	ICCV19	ResNet-50	47.5
TIN [23]	CVPR19	ResNet-50	47.8
PMFNet [40]	ICCV19	ResNet-50-FPN	48.6
†PMFNet [40]	ICCV19	ResNet-50-FPN	52.0
GPNN [1] **(Ours)**	ECCV18	ResNet-50	44.0
C-HOI$_{\text{bbox}}$ [2] **(Ours)**	CVPR20	ResNet-50	48.3
C-HOI$_{\text{mask}}$ [2] **(Ours)**	CVPR20	ResNet-50	48.9
CP-HOI$_{\text{bbox}}$ **(Ours)**	-	ResNet-50	49.9
CP-HOI$_{\text{mask}}$ **(Ours)**	-	ResNet-50	**50.4**

TABLE 4
Performance comparison on HICO-DET `test` [20] in terms of mAP$_{role}$. Higher values are better. See §4.4 for details.

Method	Default			Known Object		
	full	rare	non-rare	full	rare	non-rare
InteractNet [38]	9.94	7.16	10.77	-	-	-
Xu *et. al.* [22]	14.70	13.26	15.13	-	-	-
iCAN [43]	14.84	10.45	16.15	16.43	12.01	17.75
Wang *et. al.* [44]	16.24	11.16	17.75	17.73	12.78	19.21
TIN [23]	17.03	13.42	18.11	19.17	15.51	20.26
No-Frills [42]	17.18	12.17	18.68	-	-	-
RPNN [26]	17.35	12.78	18.71	-	-	-
PMFNet [40]	17.46	**15.65**	18.00	20.34	**17.47**	21.20
HOID [50]	17.85	12.85	19.34	-	-	-
GPNN [1] **(Ours)**	13.11	9.34	14.23	-	-	-
C-HOI [2] **(Ours)**	18.89	13.51	20.19	21.70	15.45	22.09
CP-HOI **(Ours)**	**19.42**	13.98	**20.91**	**22.01**	15.73	**22.80**

three IoU thresholds (*i.e.*, 0.25, 0.5 and 0.75). In our ablation study, we also consider R@50 and R@20 to measure the performance under stricter conditions.

Performance on HOIW: Our prior approach, C-HOI [2], reaches 1st place for relation detection on the HOIW track (*i.e.*, `test` set). The Top-5 results are listed in Table 1. As seen, the results of C-HOI are substantially better than other teams. In particular, it is **5.78%** better than the 2nd (GMVM) and **9.11%** better than the 3rd (FINet). Our approach also significantly outperforms one published state-of-the-art, *i.e.*, TIN [23]. Fig. 4 presents some visual results on HOIW `test`. Our model shows robust to various challenges, *e.g.*, occlusions, subtle relationships, *etc*.

Performance on PIC: Our C-HOI [2] also reaches 1st place for relation segmentation on the PIC `test`. We show results of the top 4 teams in Table 2. Our overall score (**52.52%**) outperforms the 2nd place by **3.85%** and the 3rd by **7.56%**. Fig. 5 depicts visual results of two complex scenes on PIC `test`. Our method shows outstanding performance in terms of instance segmentation as well as interaction recognition. It can identify both geometric and non-geometric relationships, and is capable of recognizing many fine-grained interactions, *e.g.*, *look human*, *hold tableware*. In this track, the instance localization network is instantiate as Eq. (18). In Fig. 5, we provide some visual results on PIC `test`.

4.3 Results on V-COCO

Dataset: V-COCO [19] provides verb annotations for MS-COCO[79]. Proposed in 2015, it is the first large-scale dataset for HOI understanding and remains the most popular one today. It contains 10,346 images in total ($2,533/2,867/4,946$ for `train`/`val`/`test` splits). 16,199 human instances are annotated with 26 action labels, wherein three actions (*i.e.*, *cut*, *hit*, *eat*) are annotated with two types of targets (*i.e.*, instrument and direct object), and three actions (*i.e.*, *run*, *stand*, *walk*) are annotated with no interaction object.

Evaluation Metrics: We use the original role mean AP (mAP$_{role}$), which is exactly the same as mAP$_{rel}$ in HOIW.

Performance: Since V-COCO has both bounding box and mask annotations, we provide two variants of our methods, *i.e.*, CP-HOI$_{\text{bbox}}$ and CP-HOI$_{\text{mask}}$, where CP-HOI$_{\text{bbox}}$ is trained with box annotations, while CP-HOI$_{\text{mask}}$ uses groundtruth masks. For fairness, during evaluation, the mask outputs of CP-HOI$_{\text{mask}}$ are transformed to boxes. Table 3 summarizes the results in comparison with seven state-of-the-arts. CP-HOI$_{\text{bbox}}$ outperforms TIN [23] by **2.1%** and RPNN [26] by **2.4%**. CP-HOI$_{\text{mask}}$ further improves CP-HOI$_{\text{bbox}}$ by **0.5%**, which suggests the superiority of the mask-level representation over the box-level one. We would like to note that PMFNet [40] reported a 52.0% mAP$_{role}$ on V-COCO. However, it relies on an expensive pose estimator, thus it is unfair to directly compare this with our method. Without the pose estimator, PMFNet obtains a score of 48.6%, worse than CP-HOI$_{\text{bbox}}$. In addition, compared with its preliminary versions, GPNN [1] and C-HOI [2], CP-HOI gains significant performance improvements. Some visual results on V-COCO `test` set can be found in Fig. 6.

4.4 Results on HICO-DET

Dataset: HICO-DET [20] is currently one of the largest benchmarks for HOI detection. It offers 38,118 images for `train` and 9,658 images for `test`. Each human instance is annotated with 600 HOI categories (*e.g.*, *scratching a cat*, *washing a knife*), corresponding to 80 object classes and 117 action verbs. Each image in HICO-DET has on average 1.67 instances for each HOI category.

Evaluation: Following the standard protocol [37], we conduct experiments under two different settings: 1) *Known Object* setting: for each HOI category (*e.g.*, *kicking a ball*), the performance is only evaluated on those images containing the target object category (*e.g.*, *ball*). This setting can better evaluate the accuracy in recognizing $\langle$*human, verb, object*$\rangle$ triplets; 2) *Default* setting: for each HOI category, the performance is evaluated on the full test set. This setting is more challenging since the model also needs to recognize background images (*e.g.*, images without *ball*s). In both settings, mAP$_{role}$ is used for the metric.

Performance: We compare the performance of our CP-HOI model with previous famous works in Table 4, on HICO-DET `test`. Since only bounding box level HOI annotations are provided, we only report the scores of our bounding box based model. We observe that our model performs well on the Full set of the dataset as well as on Non-rare classes. Interestingly, even though we do not target them explicitly, our model achieves competitive performance on Rare classes too. Compared with GPNN [1] and C-HOI [2],

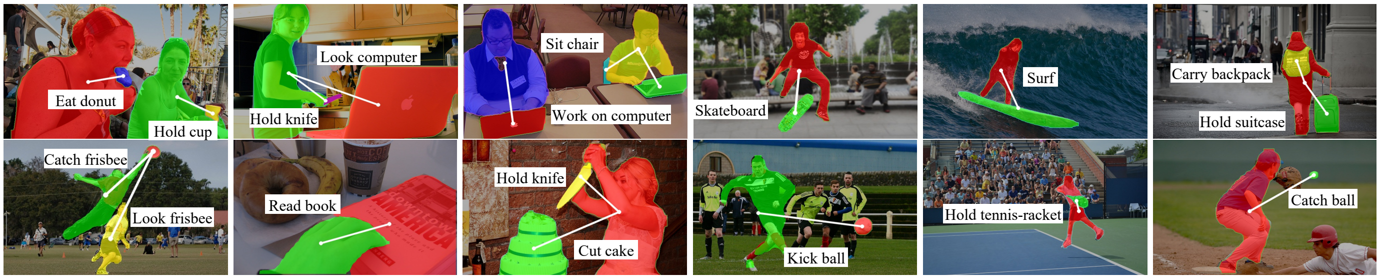


Fig. 6. Visual results for relation segmentation, on V-COCO `test` set[19].

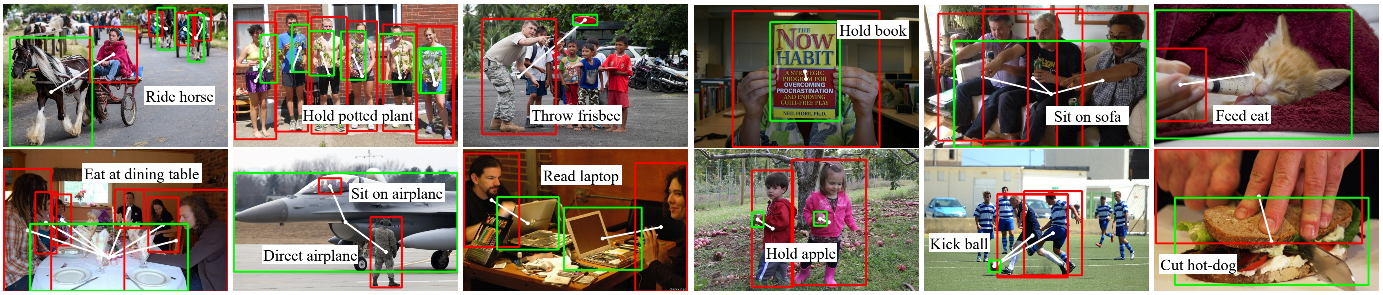


Fig. 7. Visual results for relation detection, on HICO-DET `test` set[20].

TABLE 5
Ablation study of our CP-HOI model in terms of mAP$_{role}$. Higher values are better. See §4.5 for details.

Aspect		Variant	V-COCO [19]	HICO-DET [20]					
				Default			Known Object		
				full	rare	non-rare	full	rare	non-rare
Full Model		CP-HOI ($T=3$, $S=3$)	**49.9**	**19.42**	**13.98**	**20.91**	**22.01**	**15.73**	**22.80**
Cascade Architecture		$T=1$	46.5	17.28	11.55	18.73	19.94	13.75	20.46
		$T=2$	48.5	18.30	12.82	19.90	20.99	14.60	21.78
		$T=4$	49.8	19.40	13.95	20.89	21.98	15.70	22.76
GPNN-based Relation Reasoning	Graph Structure	*w/o* graph	42.1	14.11	8.87	15.59	16.90	10.43	17.49
		w/o $\mathcal{L}_{\text{structure}}$ (Eq. (14))	46.5	17.04	11.85	19.41	19.78	13.53	20.11
		w/o $\mathcal{L}_{L_1}$ (Eq. (14))	48.5	18.33	13.05	19.12	20.81	14.76	21.29
		w/o $\mathcal{L}_{\text{rank}}$ (Eq. (16))	49.3	19.11	13.86	19.91	21.65	15.65	22.03
		w/o joint parsing	48.7	16.09	10.54	14.85	18.42	13.03	17.98
	Message Passing	$S=1$	47.9	17.91	12.40	19.83	20.55	14.34	21.10
		$S=2$	49.4	18.98	13.40	20.53	21.49	15.21	22.14
		$S=4$	49.0	19.02	13.36	20.57	21.59	15.33	22.20
		$S=5$	48.8	18.83	13.17	20.41	21.37	15.13	21.99
	Edge Embedding	*w/o* visual feature $\boldsymbol{v}$	39.7	11.61	7.84	10.21	14.51	8.97	14.31
		w/o geometric feature $\boldsymbol{g}$	46.3	15.35	10.49	16.07	18.05	12.09	19.32
		w/o semantic feature $\boldsymbol{s}$	45.8	14.79	9.98	15.42	17.93	11.67	18.39

our CP-HOI further provides considerable performance gains. In Fig. 7, we visualize some results on representative samples from HICO-DE `test` set.

4.5 Ablation Study

In this section, we analyze the contributions of different model components to the final performance and examine the effectiveness of our main assumptions. Table 5 shows the detailed results on the V-COCO and HICO-DET datasets.

Cascade Architecture. We study the impact of the number of cascade stages T used in our CP-HOI model by varying it from 1 to 4. The IoU thresholds used for these four stages are $[0.5, 0.6, 0.7, 0.75]$. As reported in Table 5, when setting $T=1$ we obtain a 46.5 mAP on V-COCO and 17.28 on HICO-DET (the Default setting). The scores are significantly improved by adding a second stage, *i.e.*, **2.0**% on V-COCO and **1.18**% on HICO-DET. When further adding more than three stages, the performance gain is marginal. Hence, considering the model complexity and performance, we choose $T=3$ as our default setting.

GPNN-based Relation Reasoning. *1) Graph structure analysis.* We first investigate the necessity of exploring structural reasoning via a graphical model for HOI recognition. We build a baseline model, *i.e.*, *w/o graph*, by directly feeding the node and edge features, which are originally used for GPNN, into a fully connected network for predicting HOI actions. It can be considered as a much simpler version of C-HOI [2]. From Table 5, we find that the performance of *w/o graph* is significantly worse than our full model over V-COCO and HICO-DET, in various settings. This demon-

TABLE 6
Comparison between mask and bbox representations (§4.5).

Relation Representation	V-COCO [19] (mAP$_{role}$)
BBox	49.9
Mask	**50.4**
BBox + Mask (`max`)	50.3
BBox + Mask (`sum`)	50.3

strates the benefits of structural modeling in this problem.

In §3.3.2, GPNN automatically infers parse graph structures via learning a soft adjacency matrix A using $\mathcal{L}_{\text{structure}}$ (Eq. (14)). To verify this strategy, we perform experiments by turning off the $\mathcal{L}_{\text{structure}}$ in the adjacency matrix learning (*w/o* $\mathcal{L}_{\text{structure}}$ in Table 5). We find that explicitly learning the graph structures provides substantial performance gain. We further examine the $\mathcal{L}_{L_1}$ and $\mathcal{L}_{\text{rank}}$ via two baseline models, *i.e.*, *w/o* $\mathcal{L}_{L_1}$ and *w/o* $\mathcal{L}_{\text{rank}}$. As can be observed, both constraints provide consistent performance improvement on V-COCO and HICO-DET.

We next study the effect of jointly learning graph structures and message passing. By isolating graph parsing from message passing, we obtain *w/o joint parsing*, where the adjacency matrices are directly computed by link functions from edge features at the beginning. We observe a performance decrease in Table 5 on both datasets, suggesting that learning graph structures and message passing together can improve the learning ability of our HOI parsing model.

2) Message passing. Now we investigate the performance with regard to the message-passing iteration step S in GPNN. We observe that setting $S = 2$ or $S = 3$ provides a substantial performance gain in mAP of $1.5\% \sim 2\%$ on V-COCO, compared to $S=1$. However, when increasing S to a certain extent (*e.g.*, $S=4$ or 5), the performance degrades slightly. This is because, with the increase of iterations, the learning ability is improved, while the risk of over-fitting is also rising. Accordingly, we chose $S=3$ for a better trade-off between accuracy and computational complexity.

3) Edge embedding. To better characterize the relations between entities, our model considers three kinds of edge features, *i.e.*, semantic feature $\boldsymbol{s}$, geometric feature $\boldsymbol{g}$ and visual feature $\boldsymbol{v}$ (Eq. (6)). Three variants, *w/o* $\boldsymbol{s}$, *w/o* $\boldsymbol{g}$, *w/o* $\boldsymbol{v}$ are accordingly built by dropping each of the features, in order to verify their effects. As seen from Table 5, *w/o* $\boldsymbol{v}$ leads to a significant performance drop (around $7\% \sim 10\%$ in terms of mAP on both datasets), verifying that the visual feature is the most important one for modeling human-object relations. Besides, the geometric and semantic features also benefit the performance. Finally, our full model considering all three features achieves the best performance.

Exploring a Better Relation Representation. Existing HOI methods typically use coarse bounding boxes represent the entities; however, is this the best choice? To answer this, we perform experiments to explore a more powerful relation representation. We evaluate the performance of our model on V-COCO `test` set using four different representations: 1) BBox; 2) Mask; 3) BBox+Mask (max); and 4) BBox+Mask (sum). Here, 1) and 2) mean that we extract the features $\boldsymbol{o}_v$ and $\boldsymbol{o}_{v,w}$ by applying RoIAlign over bbox and mask regions, respectively. 3) and 4) are the fusion of bbox and mask features with element-wise `max` and `sum` operations, respectively. Note that the detected entities are the same for all baselines. The results in Table 6 show that mask is superior to bbox, with a 0.5% improvement. The two hybrid representations are better than solely using bbox, but slightly worse than the purely mask-based. In summary, the mask-based representation does indeed benefit HOI recognition as it provides more precise information.

5 Conclusion

This paper introduces a cascaded parsing network, CP-HOI, for coarse-to-fine, structured HOI recognition. It consists of an instance detection network and an interaction recognition network, which are densely connected at each stage to fully exploit the superiority of multi-tasking. The interaction recognition network is based upon a graph parsing neural network (GPNN). GPNN is empowered with four distinct neural functions (link functions, message functions, update functions and readout functions) for iterative HOI graph inference and message-passing approximation. Our model achieves 1^{st} place on both the relation detection and relation segmentation tasks in the PIC$_{19}$ Challenge, and also outperforms prior methods on two gold standard benchmarks, V-COCO and HICO-DET. Besides, we empirically demonstrate the advantages of fine-grained masks over bounding boxes for more precise relation representations, which highlights some promising directions for future efforts.

Acknowledgment

The authors would like to thank the reviewers and editors for their time and valuable comments on this work. This work was supported by CCF-Baidu Open Fund, Zhejiang Lab's International Talent Fund for Young Professionals and Zhejiang Lab's Open Fund (No. 2019KD0AB04).

References

[1] S. Qi, W. Wang, B. Jia, J. Shen, and S.-C. Zhu, "Learning human-object interactions by graph parsing neural networks," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 401–417.

[2] T. Zhou, W. Wang, S. Qi, H. Ling, and J. Shen, "Cascaded human-object interaction recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 4263–4272.

[3] W. Norcliffe-Brown, S. Vafeias, and S. Parisot, "Learning conditioned graph structures for interpretable visual question answering," in *Proc. Advances Neural Inf. Process. Syst.*, 2018, pp. 8334–8343.

[4] L. Li, Z. Gan, Y. Cheng, and J. Liu, "Relation-aware graph attention network for visual question answering," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 10 313–10 322.

[5] Z. Zheng, W. Wang, S. Qi, and S.-C. Zhu, "Reasoning visual dialogs with structural and partial observations," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 6669–6678.

[6] W. Wang, Y. Xu, J. Shen, and S.-C. Zhu, "Attentive fashion grammar network for fashion landmark detection and clothing category classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 4271–4280.

[7] W. Wang, Z. Zhang, S. Qi, J. Shen, Y. Pang, and L. Shao, "Learning compositional neural information fusion for human parsing," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 5703–5713.

[8] H.-S. Fang, Y. Xu, W. Wang, X. Liu, and S.-C. Zhu, "Learning pose grammar to encode human body configuration for 3d pose estimation," in *The AAAI Conference on Artificial Intelligence*, 2018.

[9] J. Johnson, A. Gupta, and L. Fei-Fei, "Image generation from scene graphs," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 1219–1228.

[10] D. Shao, Y. Xiong, Y. Zhao, Q. Huang, Y. Qiao, and D. Lin, "Find and focus: Retrieve and localize video events with natural language queries," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 200–216.

[11] J. Wu, L. Wang, L. Wang, J. Guo, and G. Wu, "Learning actor relation graphs for group activity recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 9964–9974.

[12] L. Fan, W. Wang, S. Huang, X. Tang, and S.-C. Zhu, "Understanding human gaze communication by spatio-temporal graph reasoning," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 5724–5733.

[13] B. Pang, K. Zha, Y. Zhang, and C. Lu, "Further understanding videos through adverbs: A new video task," in *The AAAI Conference on Artificial Intelligence*, pp. 11 823–11 830.

[14] C.-Y. Ma, A. Kadav, I. Melvin, Z. Kira, G. AlRegib, and H. Peter Graf, "Attend and interact: Higher-order object interactions for video understanding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 6790–6800.

[15] D. Shao, Y. Zhao, B. Dai, and D. Lin, "Finegym: A hierarchical video dataset for fine-grained action understanding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 2616–2625.

[16] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi, "Optimization by simulated annealing," *Science*, vol. 220, no. 4598, pp. 671–680, 1983.

[17] P. F. Felzenszwalb, R. B. Girshick, and D. McAllester, "Cascade object detection with deformable part models," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2010, pp. 2241–2248.

[18] P. F. Felzenszwalb and D. P. Huttenlocher, "Efficient belief propagation for early vision," *Int. J. Comput. Vis.*, vol. 70, no. 1, pp. 41–54, 2006.

[19] S. Gupta and J. Malik, "Visual semantic role labeling," *arXiv preprint arXiv:1505.04474*, 2015.

[20] Y.-W. Chao, Z. Wang, Y. He, J. Wang, and J. Deng, "Hico: A benchmark for recognizing human-object interactions in images," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1017–1025.

[21] Y. Zhang, P. Tokmakov, M. Hebert, and C. Schmid, "A structured model for action detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 9975–9984.

[22] B. Xu, Y. Wong, J. Li, Q. Zhao, and M. S. Kankanhalli, "Learning to detect human-object interactions with knowledge," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 2019–2028.

[23] Y.-L. Li, S. Zhou, X. Huang, L. Xu, Z. Ma, H.-S. Fang, Y. Wang, and C. Lu, "Transferable interactiveness knowledge for human-object interaction detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 3585–3594.

[24] M. Qu and J. Tang, "Probabilistic logic neural networks for reasoning," in *Proc. Advances Neural Inf. Process. Syst.*, 2019, pp. 7710–7720.

[25] H. Tan, L. Wang, Q. Zhang, Z. Gao, N. Zheng, and G. Hua, "Object affordances graph network for action recognition," in *Proc. of the British Machine Vision Conference*, 2019.

[26] P. Zhou and M. Chi, "Relation parsing neural network for human-object interaction detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 843–851.

[27] A. Gupta and L. S. Davis, "Objects in action: An approach for combining action understanding and object perception," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2007, pp. 1–8.

[28] A. Gupta, A. Kembhavi, and L. S. Davis, "Observing human-object interactions: Using spatial and functional compatibility for recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 10, pp. 1775–1789, 2009.

[29] B. Yao and L. Fei-Fei, "Grouplet: A structured image representation for recognizing human and object interactions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2010, pp. 9–16.

[30] B. Yao, X. Jiang, A. Khosla, A. L. Lin, L. Guibas, and L. Fei-Fei, "Human action recognition by learning bases of action attributes and parts," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2011, pp. 1331–1338.

[31] V. Delaitre, J. Sivic, and I. Laptev, "Learning person-object interactions for action recognition in still images," in *Proc. Advances Neural Inf. Process. Syst.*, 2011, pp. 1503–1511.

[32] C. Desai and D. Ramanan, "Detecting actions, poses, and objects with relational phraselets," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 158–172.

[33] J.-F. Hu, W.-S. Zheng, J. Lai, S. Gong, and T. Xiang, "Recognising human-object interaction via exemplar based modelling," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2013, pp. 3144–3151.

[34] Y. Liao, S. Liu, F. Wang, Y. Chen, C. Qian, and J. Feng, "Ppdm: Parallel point detection and matching for real-time human-object interaction detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 482–490.

[35] A. Mallya and S. Lazebnik, "Learning models for actions and person-object interactions with transfer to question answering," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 414–428.

[36] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1440–1448.

[37] Y.-W. Chao, Y. Liu, X. Liu, H. Zeng, and J. Deng, "Learning to detect human-object interactions," in *Proc. IEEE Winter Conference on Applications of Computer Vision*, 2018, pp. 381–389.

[38] G. Gkioxari, R. Girshick, P. Dollár, and K. He, "Detecting and recognizing human-object interactions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 8359–8367.

[39] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Advances Neural Inf. Process. Syst.*, 2015, pp. 91–99.

[40] B. Wan, D. Zhou, Y. Liu, R. Li, and X. He, "Pose-aware multi-level feature network for human object interaction detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 9469–9478.

[41] H.-S. Fang, J. Cao, Y.-W. Tai, and C. Lu, "Pairwise body-part attention for recognizing human-object interactions," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 51–67.

[42] T. Gupta, A. Schwing, and D. Hoiem, "No-frills human-object interaction detection: Factorization, layout encodings, and training techniques," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 9677–9685.

[43] C. Gao, Y. Zou, and J.-B. Huang, "iCAN: Instance-centric attention network for human-object interaction detection," in *Proc. of the British Machine Vision Conference*, 2018.

[44] T. Wang, R. M. Anwer, M. H. Khan, F. S. Khan, Y. Pang, L. Shao, and J. Laaksonen, "Deep contextual attention for human-object interaction detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 5694–5702.

[45] J. Gu, H. Zhao, Z. Lin, S. Li, J. Cai, and M. Ling, "Scene graph generation with external knowledge and image reconstruction," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 1969–1978.

[46] K. Kato, Y. Li, and A. Gupta, "Compositional learning for human object interaction," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 234–251.

[47] B. Zhuang, Q. Wu, C. Shen, I. Reid, and A. van den Hengel, "HCVRD: a benchmark for large-scale human-centered visual relationship detection," in *The AAAI Conference on Artificial Intelligence*, 2018, pp. 7631–7638.

[48] L. Shen, S. Yeung, J. Hoffman, G. Mori, and L. Fei-Fei, "Scaling human-object interaction recognition through zero-shot learning," in *Proc. IEEE Winter Conference on Applications of Computer Vision*, 2018, pp. 1568–1576.

[49] J. Peyre, I. Laptev, C. Schmid, and J. Sivic, "Detecting unseen visual relations using analogies," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 1981–1990.

[50] S. Wang, K.-H. Yap, J. Yuan, and Y.-P. Tan, "Discovering human interactions with novel objects via zero-shot learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 11 652–11 661.

[51] L. Liu, W. Ouyang, X. Wang, P. Fieguth, J. Chen, X. Liu, and M. Pietikäinen, "Deep learning for generic object detection: A survey," *Int. J. Comput. Vis.*, vol. 128, no. 2, pp. 261–318, 2020.

[52] H. Schneiderman, "Feature-centric evaluation for efficient cascaded object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2004.

[53] P. Viola and M. J. Jones, "Robust real-time face detection," *Int. J. Comput. Vis.*, vol. 57, no. 2, pp. 137–154, 2004.

[54] Z. Cai and N. Vasconcelos, "Cascade R-CNN: Delving into high quality object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 6154–6162.

[55] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2961–2969.

[56] K. Chen, J. Pang, J. Wang, Y. Xiong, X. Li, S. Sun, W. Feng, Z. Liu, J. Shi, W. Ouyang *et al.*, "Hybrid task cascade for instance segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 4974–4983.

[57] W. Ouyang, K. Wang, X. Zhu, and X. Wang, "Chained cascade network for object detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 1938–1946.

[58] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 483–499.

[59] J. J. Tompson, A. Jain, Y. LeCun, and C. Bregler, "Joint training of a convolutional network and a graphical model for human pose estimation," in *Proc. Advances Neural Inf. Process. Syst.*, 2014, pp. 1799–1807.

[60] S. Park, B. X. Nie, and S.-C. Zhu, "Attribute and-or grammar for joint parsing of human pose, parts and attributes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 7, pp. 1555–1569, 2017.

[61] L.-C. Chen, A. Schwing, A. Yuille, and R. Urtasun, "Learning deep structured models," in *International Conference on Machine Learning*, 2015, pp. 1785–1794.

[62] F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner, and G. Monfardini, "The graph neural network model," *IEEE Trans. Neural Network*, vol. 20, no. 1, pp. 61–80, 2008.

[63] W. L. Hamilton, R. Ying, and J. Leskovec, "Representation learning on graphs: Methods and applications," *arXiv preprint arXiv:1709.05584*, 2017.

[64] J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals, and G. E. Dahl, "Neural message passing for quantum chemistry," in *International Conference on Machine Learning*, 2017, pp. 1263–1272.

[65] P. Velickovic, G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio, "Graph attention networks," in *Proc. Int. Conf. Learn. Representations*, 2018.

[66] K. Marino, R. Salakhutdinov, and A. Gupta, "The more you know: Using knowledge graphs for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 20–28.

[67] D. Xu, Y. Zhu, C. B. Choy, and L. Fei-Fei, "Scene graph generation by iterative message passing," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 5410–5419.

[68] X. Li, T. Zhou, J. Li, Y. Zhou, and Z. Zhang, "Group-wise semantic mining for weakly supervised semantic segmentation," *arXiv preprint arXiv:2012.05007*, 2020.

[69] W. Wang, X. Lu, J. Shen, D. J. Crandall, and L. Shao, "Zero-shot video object segmentation via attentive graph neural networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 9236–9245.

[70] J. Yang, J. Lu, S. Lee, D. Batra, and D. Parikh, "Graph R-CNN for scene graph generation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 670–685.

[71] X. Lu, W. Wang, M. Danelljan, T. Zhou, J. Shen, and L. Van Gool, "Video object segmentation with episodic graph memory networks," in *Proc. Eur. Conf. Comput. Vis.*, 2020.

[72] W. Wang, H. Zhu, J. Dai, Y. Pang, J. Shen, and L. Shao, "Hierarchical human parsing with typed part-relation reasoning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 8929–8939.

[73] O. Ulutan, A. Iftekhar, and B. S. Manjunath, "Vsgnet: Spatial attention network for detecting human object interactions using graph convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 13 617–13 626.

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

[75] R. Zellers, M. Yatskar, S. Thomson, and Y. Choi, "Neural Motifs: Scene graph parsing with global context," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 5831–5840.

[76] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[77] K. Cho, B. Van Merriënboer, D. Bahdanau, and Y. Bengio, "On the properties of neural machine translation: Encoder–decoder approaches," *Syntax, Semantics and Structure in Statistical Translation*, p. 103, 2014.

[78] Y. Li, D. Tarlow, M. Brockschmidt, and R. Zemel, "Gated graph sequence neural networks," *Proc. Int. Conf. Learn. Representations*, 2016.

[79] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 740–755.

Tianfei Zhou received his Ph.D. degree from Beijing Institute of Technology in 2017. He is currently a postdoc researcher at ETH Zurich, Switzerland. From 2019 to 2020, he was a research associate at Inception Institute of Artificial Intelligence (IIAI), Abu Dhabi, United Arab Emirates. His current research interests include human-object interaction recognition, video object segmentation and deep learning.

Siyuan Qi received his B.Eng. degree in Computer Engineering from the University of Hong Kong in 2013. He received his M.S. and Ph.D degree in Computer Science from University of California, Los Angeles in 2015 and 2019, respectively. He is currently at Google, USA. His research interests include pattern recognition, machine learning and computer vision, with a focus on human activity recognition, scene understanding with compositional representations.

Wenguan Wang received his Ph.D. degree from Beijing Institute of Technology in 2018. He is currently a postdoc researcher at ETH Zurich, Switzerland. From 2016 to 2018, he was a visiting Ph.D. student in University of California, Los Angeles. From 2018 to 2019, he was a senior scientist at Inception Institute of Artificial Intelligence, UAE. His current research interests include computer vision, image processing and deep learning.

Jianbing Shen (M'11-SM'12) is currently acting as the Lead Scientist with the Inception Institute of Artificial Intelligence, Abu Dhabi, United Arab Emirates. He is also a Full Professor with the School of Computer Science, Beijing Institute of Technology. He has published about 100 journal and conference papers such as *IEEE TPAMI*, *CVPR*, and *ICCV*. He has obtained many honors including the Fok Ying Tung Education Foundation from Ministry of Education, the Program for Beijing Excellent Youth Talents from Beijing Municipal Education Commission, and the Program for New Century Excellent Talents from Ministry of Education. His research interests include computer vision and deep learning. He is an Associate Editor of *IEEE TIP*, *IEEE TNNLS* and other journals.

Song-Chun Zhu received Ph.D. degree from Harvard University in 1996, and is Chair Professor jointly with Tsinghua University and Peking University, director of Institute for Artificial Intelligence, Peking University. He worked at Brown, Stanford, Ohio State, and UCLA before returning to China in 2020 to launch a non-profit organization – Beijing Institute for General Artificial Intelligence. He has published over 300 papers in computer vision, statistical modeling and learning, cognition, Language, robotics, and AI. He received a number of honors, including the Marr Prize in 2003, the Aggarwal prize from the Intl Association of Pattern Recognition in 2008, the Helmholtz Test-of-Time prize in 2013, twice Marr Prize honorary nominations in 1999 and 2007, a Sloan Fellowship, the US NSF Career Award, and ONR Young Investigator Award in 2001. He is a Fellow of IEEE since 2011. He serves as General co-Chair for CVPR 2012 and CVPR 2019.