Learning Human-Object Interaction via Interactive Semantic Reasoning

Dongming Yang, Yuexian Zou, *Senior Member, IEEE*, Zhu Li, *Senior Member, IEEE*, and Ge Li, *Senior Member, IEEE*

***Abstract*— Human-Object Interaction (HOI) detection devotes to learn how humans interact with surrounding objects via inferring triplets of ⟨ human, verb, object ⟩. Recent HOI detection methods infer HOIs by directly extracting appearance features and spatial configuration from related visual targets of human and object, but neglect powerful interactive semantic reasoning between these targets. Meanwhile, existing spatial encodings of visual targets have been simply concatenated to appearance features, which is unable to dynamically promote the visual feature learning. To solve these problems, we first present a novel semantic-based Interactive Reasoning Block, in which interactive semantics implied among visual targets are efficiently exploited. Beyond inferring HOIs using discrete instance features, we then design a HOI Inferring Structure to parse pairwise interactive semantics among visual targets in scene-wide level and instance-wide level. Furthermore, we propose a Spatial Guidance Model based on the location of human body-parts and object, which serves as a geometric guidance to dynamically enhance the visual feature learning. Based on the above modules, we construct a framework named Interactive-Net for HOI detection, which is fully differentiable and end-to-end trainable. 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 5.9% and 17.7% relatively, validating its efficacy in detecting HOIs.**

***Index Terms*— Human-object interaction, interactive reasoning, spatial encoding, human body-part.**

I. 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 [1], [2], segmentation [3], [4] and human pose estimation [5], [6], detecting HOIs requires a deeper understanding of visual semantics to depict complex relationships between ⟨ human, object ⟩ pairs. HOI detection is related to action recognition [7]–[9] 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. HOI detection can be widely used in applications like intelligent monitoring and man-machine interaction.

Most existing HOI detection approaches infer HOIs by employing the appearance features [10] and spatial configuration [11] of a person and an object extracted from Convolutional Neural Networks (CNNs) [12], [13], which may neglect the detailed and high-level interactive semantics implied between the related targets. To remedy the limitation above, a number of algorithms employ additional contextual cues from the image such as human intention [14] and attention boxes [15]. More recently, several works adopt human pose [14], [16] and body-parts [17], [18], or explore spatial encodings [19] to infer HOIs. Although directly incorporating contextual cues 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 methods are limited by treating visual targets separately without considering crucial semantic dependencies among them (i.e., scene, human and object).
- Secondly, concatenating human pose features or spatial encodings directly is insufficient to exploit the spatial relation between human and object. For instance, the fine-grained geometric distribution of human body-parts and object could promote the visual feature learning as an auxiliary spatial indicator.

To address issues above, we come up with the innovation of interactive semantic reasoning, which exploits HOI-specific representation by connecting relational visual targets in different level for reasoning. As illustrated in Figure 1, our model goes beyond current approaches that lacks the capability of reasoning interactive semantics. In addition, we creatively exploit the fine-grained geometric distribution from visual targets (i.e., human and object) and build a learnable module

Manuscript received June 10, 2020; revised June 9, 2021; accepted October 23, 2021. Date of publication November 9, 2021; date of current version November 12, 2021. This work was supported in part by the National Engineering Laboratory for Video Technology-Shenzhen Division and in part by the Shenzhen Municipal Development and Reform Commission (Disciplinary Development Program for Data Science and Intelligent Computing). The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Riccardo Leonardi. *(Corresponding author: Yuexian Zou.)*

Dongming Yang is with the School of Electronic and Computer Engineering (ECE), Peking University, Shenzhen 518055, China (e-mail: yangdongming@pku.edu.cn).

Yuexian Zou is with the School of Electronic and Computer Engineering (ECE), Peking University, Shenzhen 518055, China, and also with the Peng Cheng Laboratory, Shenzhen 518055, China (e-mail: zouyx@pku.edu.cn).

Zhu Li is with the Department of Computer Science and Electrical Engineering, University of Missouri–Kansas City, Kansas City, MO 64110 USA (e-mail: lizhu@umkc.edu).

Ge Li is with the School of Electronic and Computer Engineering (ECE), Peking University, Shenzhen 518055, China, and also with the Peng Cheng Laboratory, Shenzhen 518055, China (e-mail: gli@pkusz.edu.cn).

Digital Object Identifier 10.1109/TIP.2021.3125258

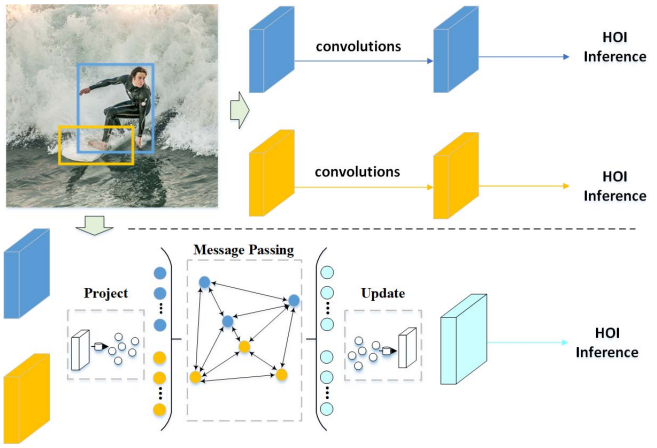

Fig. 1. A simplified schematic of how our model facilitates HOI detection. The baseline framework (above) adopts appearance features discretely to infer HOIs and neglects interactive reasoning. We present a novel model (bottom) that reasons interactive semantics among visual targets to learn HOI-specific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

to promote visual feature learning. The main contributions of this work are fourfold:

- A novel semantic-based Interactive Reasoning Block (abbr. IRB) is proposed, which reasons and integrates strong interactive semantics among visual targets. In particular, IRB 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 feature representation.
- Secondly, based on the proposed IRB, a HOI Inferring Structure is designed to implicitly parse interactive semantics in two levels (i.e., scene-wide level and instance-wide level) for inferring HOIs. Differing from treating each visual target separately, HOI Inferring Structure enables the framework to reason pairwise interactive semantics for HOI-specific modeling.
- Moreover, in order to exploit fine-grained spatial information and make full use of the geometric distribution between human and object, a Spatial Guidance Model (abbr. SGM) based on the location of human body-parts and object is proposed, which serves as the geometric guidance to promote the visual feature learning for better HOI recognition.
- Building on the modules above, we offer a general framework referred to as Interactive-Net for HOI detection. The proposed Interactive-Net is a two-stream network, in which the final HOI predictions are made by combining all exploited semantics.

We perform extensive experiments on two public benchmarks, i.e., V-COCO [20] dataset and HICO-DET [21] dataset. Our method provides obvious performance gain compared with the baseline and outperforms the state-of-the-art methods by a sizable margin. We also provide detailed ablation studies of our method to facilitate the future research.

This paper is partly based on our preliminary conference work [22] and we extend it in a number of significant aspects. First, we bring a more efficient instantiation of interactive semantic reasoning (i.e., IRB), and introduce the HOI Inferring Structure, which is innovative in theory and algorithm (e.g., the Interactive Body Attention Mechanism). Second, we add more exploration on the geometric distribution of human and object, bringing a newly proposed Spatial Guidance Model. Third, we extend the original pipeline and bring more additional analysis and experimental validations. Finally, our newly proposed Interactive-Net achieves novel performance enhancements.

II. Related Work

A. Contextual Cues in HOI Detection

The early human activity recognition [23], [24] task is confined to scenes containing single human-centric action and ignores spatial localization of the person and related object. Therefore, Gupta [20] introduced visual semantic role labeling to learn interactions between human and object. HO-RCNN [21] 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 to improve HOI detection. Auxiliary boxes [25] were employed to encode context regions from the human bounding boxes. InteractNet [10] extended the object detector Faster R-CNN [1] with an additional branch and estimated an action-specific density map to identify the locations of interacted objects. iHOI [14] utilized human gaze to guide the attended contextual regions in a weakly-supervised setting for learning HOIs. Beyond all that, spatial configuration and human pose have become the most popular cues.

1) Spatial Configuration: Given a ⟨ human, object ⟩ pair with candidate boxes, the earliest interaction pattern [21] or spatial configuration [11], [26] was a coarse layout encoded by a binary image with two channels, where the first channel has value 1 at pixels enclosed by the human bounding box, and value 0 elsewhere; the second channel has value 1 at pixels enclosed by the object bounding box, and value 0 elsewhere. TIN [16] further concatenated the above spatial configuration with human pose to encode a spatial-pose map, in which the human keypoints with lines of different gray value ranging from 0.15 to 0.95 are linked to represent different body-parts. Analogously, Tanmay [19] explored both a coarse layout of human-object box-pair and a fine-grained layout of human pose, in which the absolute position and relative configuration of the human and object boxes, as well as the absolute human pose and the relative location with respect to the object candidate box were encoded. Although concatenating the spatial encodings above helped to eliminate certain predictions and improve the detection results, these methods neglected to explore fine-grained geometric distribution as a guidance to dynamically improve the visual feature learning.

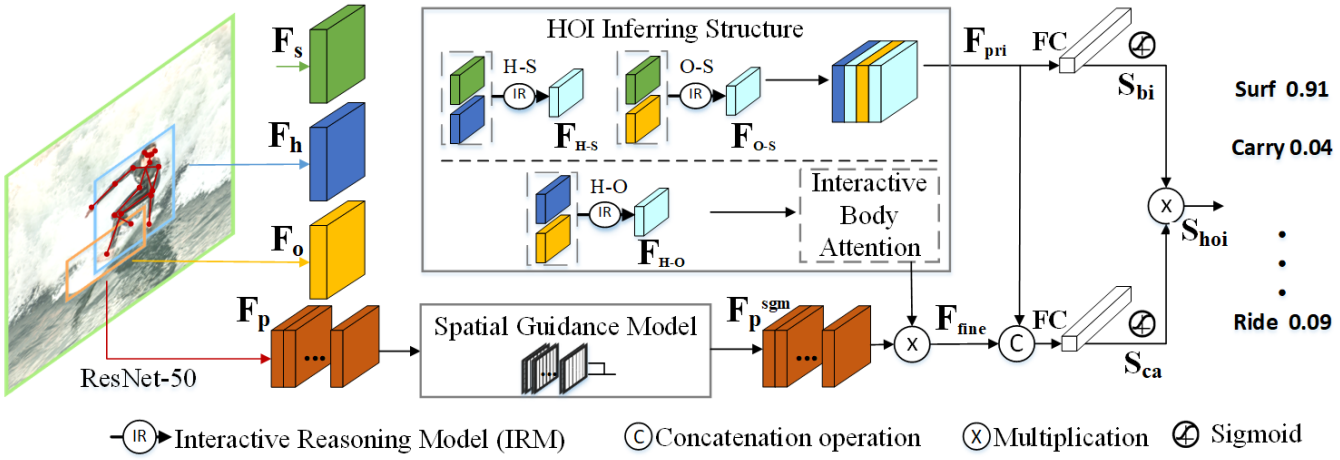

Fig. 2. An overview of Interactive-Net. The framework consists of two branches, which are a primary branch and a fine-grained branch. In the HOI Inferring Structure, we employ three IRBs and integrate them into two-level to infer HOIs. Meanwhile, our proposed Spatial Guidance Model is built into the human body-part features to enhance its representation.

2) Human Pose: Very recently, human pose [16]–[18], [27] obtained from pre-trained human pose estimators [2], [5], [28] has been widely adopted as an significant cue to tackle HOI detection. Pair-wise human parts correlation [27] was exploited to learn HOIs. PMFNet [18] developed a multi-branch network to learn a pose-augmented relation representation to incorporate interaction context, object features and detailed human parts. RPNN [17] 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 human pose benefits feature expression, there is still room for improvement. For instance, beyond employing all pose features directly, it is valuable to assign appropriate weights to different body-parts (e.g., some occluded or even invisible body-parts should be given low weights) according to the human-object interactive semantics. Besides, utilizing the geometric distribution of human body-parts to improve the visual feature learning is worth exploring.

B. Semantic Reasoning

1) Attention Mechanism: Attention mechanism [7] in action recognition helped to suppress irrelevant global information and highlight informative regions. Inspired by action recognition methods [29], [30], iCAN [11] and GID-Net [31] exploited an instance-centric attention mechanism to enhance the information from regions and facilitate HOI classification. Furthermore, Contextual Attention [32] proposed a deep contextual attention framework for HOI detection, in which context-aware appearance features for human and object were captured. PMFNet [18] focused on pose-aware attention for HOI detection by employing attentional human parts. VSGNet [26] proposed a spatial attention branch, which refined visual features in a coarse way by the spatial configuration of human-object pair. Overall, methods with attention mechanisms learned 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.

2) Semantic-Based Reasoning: Graph Parsing Neural Network (GPNN) [33] introduced a learnable semantic-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 [34], using a message function and a Gated Recurrent Unit (GRU) [35] to update states iteratively. 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 was unable to handle complex interactions. In addition, it required iteratively message passing and updating. Thirdly, it excluded semantic information from the scene for inferring HOIs. Lately, RPNN [17] introduced a structure with two semantic-based modules incorporated together to infer HOIs, which is enlightening but complicated and had heavy computing burden. In this paper, we aim to develop a novel semantic-based model to provide interactive semantic reasoning between visual targets in multiple level. Instead of coarse and iterative message passing between instances, our model captures pixel-level interactive semantics between targets all at once.

III. Proposed Method

A. Overview of Interactive-Net

1) General Framework: An overview of our proposed Interactive-Net is shown in Figure 2. ResNet-50 [13] is employed as the backbone network in our implement to extract the shared global convolutional feature map, after which the RoI-Align [2] with resolution of 7*7 according to respective candidate boxes (i.e., scene boxes, human boxes, object boxes) outputs the features for visual targets T_{scene}, T_{human}, and T_{object}, teamed as F_s, F_h and F_o. The human body-part features F_p are also extracted from the shared global convolutional feature map according to the human keypoint locations (See Section III-C).

Based on our proposed Interactive Reasoning Blocks (abbr. IRBs, See Section III-B), the HOI Inferring Structure is able

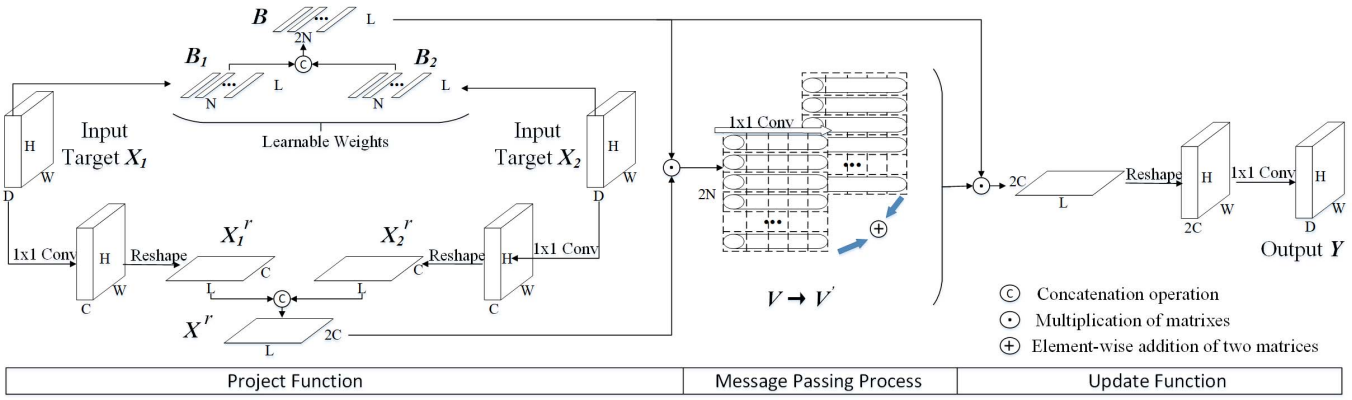

Fig. 3. The detailed design of proposed Interactive Reasoning Block (IRB), which is a more efficient instantiation of in-Graph [22]. The model consists of three core procedures, which are a project function, a message passing process and an update function. The IRB takes convolutional features of two 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.

to reason the interactive semantics for HOI-specific feature representation. In the proposed structure, visual targets T_{scene}, T_{human}, and T_{object} are assigned into three IRBs in pairs. Concretely, the employment of above IRBs can be divided into two levels (i.e., scene-wide level and instance-wide level). The scene-wide reasoning contains a human-scene (H-S) IRB and an object-scene (O-S) IRB; The instance-wide reasoning refers to the human-object (H-O) IRB, which been embedded into the human body-parts via an Interactive Body Attention Mechanism (abbr. IBAM).

Moreover, our proposed Spatial Guidance Model (abbr. SGM, See Section III-D) is built into the human body-part features to improve its ability to differentiate interactions with diverse geometric distributions.

2) Two Branches: Sharing the similar inspiration with TIN [16], Interactive-Net is a two-stream network containing a primary branch and a fine-grained branch. In the primary branch, feature F_{pri} is fed into fully connected layers to predict whether any interaction exists in the ⟨ human, object ⟩ candidate. This makes the HOI detector suppress most non-interactive ⟨ human, object ⟩ candidates and improve the final detection. In the fine-grained branch, feature F_{pri} and feature F_{fine} are concatenated and fed into fully connected layers to predict the HOI category labels for the ⟨ human, object ⟩ candidate. As a person is able to concurrently perform different interactions with one or multiple objects, each prediction of an interaction is independent and not mutually exclusive. Therefore, for a HOI candidate, Interactive-Net concurrently predicts the interactiveness label (i.e., "interactive" or "non-interactive") with score S_{bi} and HOI category labels with scores $S_{cate} = [S^1_{cate}, S^2_{cate}, \ldots, S^C_{cate}]$ by binary sigmoid classifiers. Here, C denotes the number of HOI categories.

$$S_{bi} = sigmoid(Fc(F_{pri})), \tag{1}$$

$$S^c_{cate} = sigmoid(Fc(F_{pri} \oplus F_{fine})), \tag{2}$$

where $\oplus$ refers to the concatenation operation and $c \in \{1, \ldots, C\}$ denotes the HOI category. Similar to general HOI detection frameworks [10], [11], we use a fusion of scores obtained from branches to predict final scores S_{hoi} for each HOI:

$$S^c_{hoi} = S_{bi} * S^c_{cate}. \tag{3}$$

In each image, pairwise candidate boxes ($B_h * B_o$) for each HOI category are finally assigned binary labels based on the prediction. Since HOI detection is a multi-label classification problem, our Interactive-Net is trained in a supervised fashion using multi-label binary cross-entropy loss $L_{CE}(\cdot)$. Two branches are trained jointly, where the overall loss is the sum of losses from two branches.

$$L_{overall} = \frac{1}{N}\sum_{n=1}^{N}[L_{CE}(g_{bi}, S_{bi}) + \sum_{c=1}^{C} L_{CE}(g^c_{cate}, S^c_{cate})], \tag{4}$$

where g_{bi} and $g^c_{cate} \in G_{cate}$ denote the ground truth labels of interactiveness and HOI categories, respectively. N is the size of training set. In this way, the entire framework is implemented to be fully differentiable and end-to-end trainable using gradient-based optimization.

B. Interactive Reasoning Block

The detailed design of proposed IRB is provided in Figure 3. 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 IRB 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 as input. Here $H*W$ denotes range and D denotes feature dimension. We first propose a project function to map two feature tensors into a semantic-based interactive space, where a fully-connected graph structure can be built. Based on the graph structure, message passing process is then employed 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 outputs feature Y, enabling us to utilize the reasoned semantics in convolution space. We then describe its architecture in details.

1) Project Function: Project function aims to provide an pattern $P(\cdot)$ to fuse two targets together, after which message passing process can be efficiently computed. The calculation process of $P(\cdot)$ 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.

a) 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 dimension C, obtaining $X^r = [X_1^r, X_2^r] \in \Re^{L*2C}$.

b) 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.

c) 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 $x_j^r \in \Re^{1*2C}$, $v_i \in \Re^{1*2C}$.

$$v_i = \sum_{\forall j} b_{ij} x_j^r. \tag{5}$$

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.

2) Message Passing Process: 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 structure, message passing process is adopted to broadcast and integrate semantic information from all nodes over the graph.

GPNN [33] applies an iterative process with GRU [35] 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$M(\cdot)$ is computed by:

$$M(V) = \omega(AV) = Conv1D(V) \oplus V. \tag{6}$$

Here A denotes the adjacency matrix among nodes learned by gradient decent during training, reflecting the weights for edge E. The $\omega(\cdot)$ 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 [36] 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.

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

$$y_i^r = \sum_{\forall j} b_{ij} v_j', \tag{7}$$

here $v_j' \in \Re^{1*2C}$, $y_i^r \in \Re^{1*2C}$.

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. The IRBs have input dimension $D = 256$, reduced dimension $C = 128$, number of nodes $N = 64$ according to our empirical tests (See Section IV-C), which is a more efficient instantiation of in-Graph model [22] to be integrated into the network. In this way, updated features in convolution space can play its due role in the following schedule.

C. HOI Inferring Structure

Our IRB 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 ⟨ human, surf, surfboard ⟩ 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 IRBs in two levels to build a HOI Inferring Structure, which are scene-wide level and instance-wide level. The scene-wide reasoning contains a human-scene (H-S) IRB and an object-scene (O-S) IRB, the instance-wide reasoning refers to the human-object (H-O) IRB.

In scene-wide reasoning, semantic features F_{H-S} and F_{O-S} obtained from H-S IRB and O-S IRB are concatenated to F_h and F_o respectively to enrich feature representation.

In instance-wide reasoning, features F_{H-O} obtained from H-O IRB is further embedded into human body-parts through an Interactive Body Attention Mechanism (abbr. IBAM).

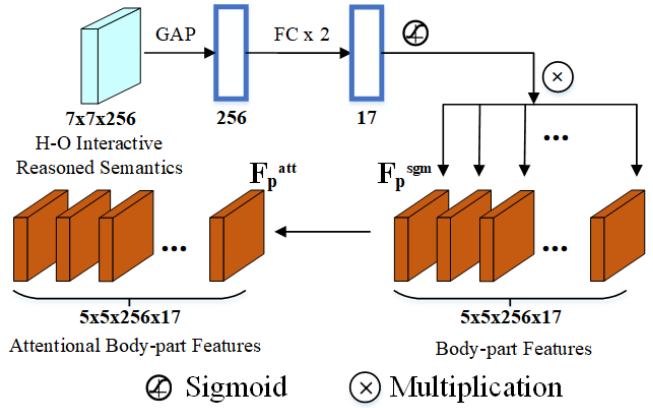

Fig. 4. A visualization of Interactive Body Attention Mechanism. We encode and transmit the H-O interactive semantics into human body-parts to learn the importance of each body-part to the interaction.

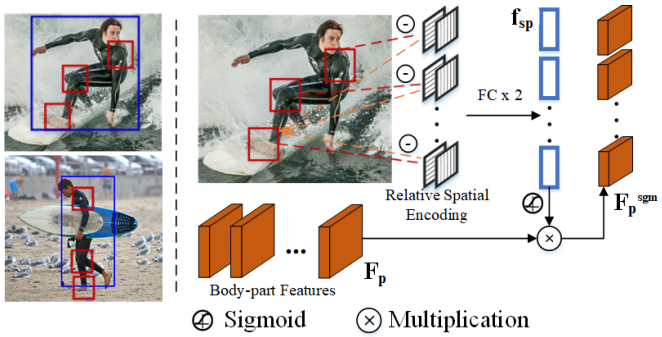

Fig. 5. Spatial Guidance Model (SGM) visualization. The left side shows two HOI examples (i.e., ⟨ human, surf, surfboard ⟩, ⟨ human, hold, surfboard ⟩) where human body-parts have distinct spatial distribution relative to the object (i.e. surfboard). The right side visualizes our proposed SGM which exploits the geometric distribution of human body-parts and object to promote visual feature learning.

Given the human keypoint locations $K = [k_1, k_2, \ldots, k_n]$ estimated by the pose estimator [28], where $n = 17$ and $k_i = [k_i^x, k_i^y]$ donating the coordinates, we first extract fine-grained human body-part bounding boxes $P = [p_1, p_2, \ldots, p_n]$ by crop local regions according to human keypoint locations on the shared global convolutional feature map. Specifically, for every body-part, the local bounding box $p_i = [x_i^1, y_i^1, x_i^2, y_i^2]$ has the calculation as follow:

$$\begin{aligned} x_i^1 &= k_i^x - \alpha * w_h, \\ y_i^1 &= k_i^y - \alpha * h_h, \\ x_i^2 &= k_i^x + \alpha * w_h, \\ y_i^2 &= k_i^y + \alpha * h_h, \end{aligned} \tag{8}$$

where $x_i^1, y_i^1, x_i^2, y_i^2$ denote the 2D coordinates of bounding box p_i, (w_h, h_h) is the size of human bounding box and $\alpha = 0.1$ controls the body-part size according to (w_h, h_h). Then, we adopt RoI-Align [2] with size of $5 * 5$ to rescale the body-part regions, getting the pooled body-part features $F_p = [f_{p1}, f_{p2}, \ldots, f_{pn}]$.

It is obvious that body-parts typically have strong correlations with interactions and each body-part assumes different importance for recognizing an interaction. Besides, some body-parts may be blocked or even invisible sometimes. Therefore, we propose an Interactive Body Attention Mechanism to transmit H-O interactive semantics into human body-parts and learn the importance of each body-part to the interaction. As illustrated in Figure 4, a global average pooling (GAP) and a multilayer perceptron (MLP) with two fully connected layers followed by a sigmoid operation are adopted on feature F_{H-O}, generating the embedded vector $\beta = [\beta_1, \beta_2, \ldots, \beta_n]$ as the semantic attention value. Then body-part features F_p^{sgm} output from Spatial Guidance Model (abbr. SGM, See Section III-D) are weighted by a multiplication operation, generating the attentional body-part features F_p^{att}:

$$\beta = sigmoid(Fc(GAP(F_{H-O}))), \tag{9}$$

$$F_p^{att} = \beta * F_p^{sgm}, \tag{10}$$

where $\beta_i \in [0, 1]$ is the $i - th$ semantic attention value for f_{pi}^{sgm}.

In this way, compared with in-GraphNet [22], the HOI Inferring Structure here provides interactive semantic reasoning not only for primary branch but also for fine-grained branch to explore importance of human body-parts. Therefore, it can help to learn more discriminative and fine-grained representation to recognize diverse HOIs.

D. Spatial Guidance Model

As shown in Figure 5, we present the Spatial Guidance Model in this subsection, which is able to employ the geometric distribution of human body-parts and object to dynamically promote the visual feature learning. Differing from the coarse and absolute layout presented by the two-channel binary image [11], [21] or the complicated spatial configuration [19], we simply encode relative spatial relationship between human and object using human keypoints and object bounding box.

We first generate a two-channel coordinate map for the shared global convolutional feature map, where the first channel indicates the x coordinate and the second channel indicates the y coordinate for each pixel. Then, we encode the relative location of each pixel to the object by calculating spatial offset of (x, y) coordinates relative to object center $(x_{o\_center}, y_{o\_center})$. After that, the relative spatial encodings for human body-parts can be obtained by employing the same crop and RoI-Align operation (See Section III-C) on the normalized coordinate maps, which match the human body-part bounding boxes $P = [p_1, p_2, \ldots, p_n]$.

We further adopt a MLP with two fully connected layers and a sigmoid operation to extract the feature f_{sp} from the relative spatial encoding. Finally the output feature f_{sp}, as a significant spatial indicator, promotes body-part feature learning by an element-wise multiplication.

$$F_p^{sgm} = f_{sp} * F_p, \tag{11}$$

With formulations above, rich interactive semantics and geometric relations among visual targets can be explicitly utilized to infer HOIs.

IV. Experiments and Evaluations

A. Experimental Setup

1) Datasets and Evaluation Metrics: We evaluate our method and compare it with the state-of-the-arts on two large-scale benchmarks, including V-COCO [20] and HICO-DET [21] datasets. V-COCO [20] includes 10,346 images, which is a subset of MS COCO dataset [37]. It contains 16,199 human instances in total and provides 26 common HOI annotations. Each person in V-COCO is annotated with a binary label for each HOI category, indicating whether the person is performing the certain interaction. Thus, each person can perform multiple interactions at the same time. HICO-DET [21] contains about 48k images and 600 HOI categories over 80 object categories, which provides more than 150K annotated ⟨ human, object ⟩ pairs. There are three different HOI category sets in HICO-DET, which are: (a) all 600 HOI categories (Full), (b) 138 HOI categories with less than 10 training instances (Rare), and (c) 462 HOI categories with 10 or more training instances (Non-Rare). There also exist two different evaluation settings: (a) Default setting: all images both containing and not containing the target object category are evaluated. (b) Known Object setting: only images containing the target object category are evaluated.

A detected ⟨ human, verb, object ⟩ triplet is considered as a true positive if 1) it has the correct interaction label and 2) both the predicted human and object bounding boxes have IoU of 0.5 or higher with the ground-truth boxes. We use role mean average precision (role mAP) [20] on both benchmarks.

2) Implementation Details: Following the protocol in [11], human and object bounding boxes are generated using the ResNet-50 version of Faster R-CNN [1]. Human boxes with scores higher than 0.8 and object boxes with scores higher than 0.4 are kept for detecting HOIs. The pose estimator CPN [28] has been used to provide $K = 17$ keypoints for each human candidate. We train our model with Stochastic Gradient Descent (SGD), using a learning rate of 1e-3, a weight decay of 1e-4, and a momentum of 0.9. The model is trained for 50K and 300K iterations on V-COCO and HICO-DET, respectively.

B. Overall Performance

We compare our method with several state-of-the-arts in this subsection. Meanwhile, we strip all modules related to Interactive Reasoning Blocks, HOI Inferring Structure and Spatial Guidance Model from our proposed framework as the baseline.

1) Performance on V-COCO: Comparison results on V-COCO in terms of role mAP are shown in Table I. It can be seen that our proposed Interactive-Net has a mAP of 52.1, obtaining the best performance among all methods. Although we do not adopt a separate spatial configuration branch as TIN [16] and VSGNet [26] do, our method outperforms these methods with sizable gains. Besides, our method achieves an absolute gain of 2.9 points compared with the baseline, which is a relative improvement of 5.9%, validating its efficacy in HOI detection task.

TABLE I
Performance Comparison With State-of-the-Arts on V-COCO Dataset

Method	Backbone Network	mAP$_{role}$(%)
Gupta et al. [20]	ResNet-50-FPN	31.8
InteractNet [10]	ResNet-50-FPN	40.0
iHOI [14]	ResNet-50	40.4
BAR-CNN [15]	Inception-ResNet	41.1
GPNN [33]	Deformable CNN	44.0
iCAN [11]	ResNet-50	45.3
Contextual Att [32]	ResNet-50	47.3
RPNN [17]	ResNet-50	47.5
TIN(RP_dC_d) [16]	ResNet-50	47.8
VSGNet [26]	ResNet-50	<u>51.0</u>
our baseline	ResNet-50	49.2
Interactive-Net (ours)	ResNet-50	**52.1**

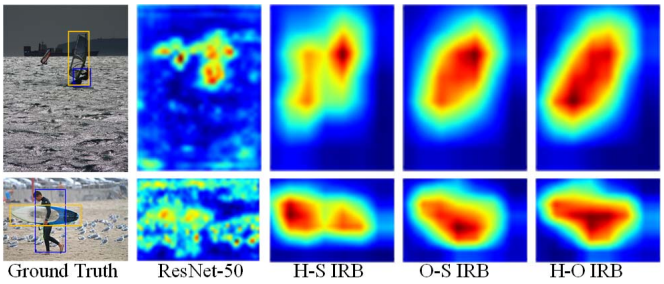

Fig. 6. (a) Interactive Reasoning Block visualization. The first column shows two HOI examples (i.e., ⟨ human, surf, surfboard ⟩, ⟨ human, hold, surfboard ⟩). The rest of columns visualize feature responses from ResNet-50, H-S IRB, O-S IRB and H-O IRB, respectively.

2) Performance on HICO-DET: Table II shows the comparisons of Interactive-Net and state-of-the-arts on HICO-DET. We report the quantitative evaluation of Full, Rare, and Non-rare HOIs with two different settings: Default and Known Object. Firstly, the detection results of our Interactive-Net are higher than the second best under all evaluation settings, demonstrating that our method is more competitive than the others in detecting all kinds of HOIs. Besides, our Interactive-Net obtains 18.21 mAP on HICO-DET (Default Full mode), which achieves absolute gains of 2.7 points and relative gains of 17.7% compared with the baseline. These results quantitatively show the efficacy of our method.

3) Time Consumption: While above state-of-the-arts have not report their detection speed, we compare our method with the baseline in this subsection. We run Interactive-Net and the baseline on V-COCO testset under the same hardware and software conditions. The comparison shows that our Interactive-Net processes 1.3 images per second during testing, which is very close to the baseline of 1.5 images per second. This means that, our proposed modules improve the performance effectively but bring very small time consumption.

TABLE II

COMPARISON RESULTS ON HICO-DET TEST SET. DEFAULT: ALL IMAGES. KNOWN OBJECT: ONLY IMAGES CONTAINING TARGET OBJECT CATEGORY. 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

Method	Backbone Network	Default			Know Object		
		Full	Rare	Non-rare	Full	Rare	Non-rare
Shen et al. [38]	VGG-19	6.46	4.24	7.12	-	-	-
HO-RCNN [21]	CaffeNet	7.81	5.37	8.54	10.41	8.94	10.85
InteractNet [10]	ResNet-50-FPN	9.94	7.16	10.77	-	-	-
iHOI [14]	ResNet-50	9.97	7.11	10.83	-	-	-
GPNN [33]	Deformable ConvNet	13.11	9.34	14.23	-	-	-
iCAN [11]	ResNet-50	14.84	10.45	16.15	16.26	11.33	17.73
Contextual Att [32]	ResNet-50	16.24	11.16	17.75	17.73	12.78	19.21
RPNN [17]	ResNet-50	17.35	12.78	18.71	-	-	-
TIN(RP_dC_d) [16]	ResNet-50	17.03	13.42	18.11	19.17	15.51	20.26
No-Frills [19]	ResNet-152	17.18	12.17	18.68	-	-	-
our baseline	ResNet-50	15.47	11.75	16.58	18.81	15.06	19.93
Interactive-Net (ours)	ResNet-50	**18.21**	**14.19**	**19.41**	**20.74**	**16.67**	**21.96**

TABLE III

IMPACT OF HOI INFERRING STRUCTURE

H-S IRB		√		√		√
O-S IRB			√	√		√
H-O IRB with IBAM					√	√
mAP$_{role}$(%)	49.2	50.2	50.3	50.5	50.6	50.9

TABLE IV

BUILDING INTERACTIVE SPACES WITH DIFFERENT NUMBER OF NODES

number of nodes (N)	16	32	64	128	256
mAP$_{role}$(%)	51.7	51.8	52.1	51.9	51.6

TABLE V

IMPACT OF ADOPTING INTERACTIVE BODY ATTENTION MECHANISM. HERE, *concat*() MEANS FEATURE CONCATENATION

Method	mAP$_{role}$(%)
baseline	49.2
baseline+$concat(F_{H-O}, F_{pri})$	49.9
baseline+$concat(F_{H-O}, F_p^{sgm})$	49.4
baseline+(F_{H-O} with IBAM	50.6

C. Ablation Studies

We conduct several ablation studies in this subsection. V-COCO serves as the primary testbed on which we further analyze the individual effect of components in our method.

1) Performance Impact of Proposed Components: In this subsection, we study the performance impact of HOI Inferring Structure and Spatial Guidance Model. As shown in Table III, we perform an ablation study on HOI Inferring Structure by testing three IRBs severally. When we only adopt the H-S IRB, the O-S IRB or the H-O IRB with IBAM, the mAP are 50.2, 50.3 and 50.6 respectively, indicating the respective effects of these three blocks. H-S IRB and O-S IRB together compose the scene-wide reasoning, obtaining the mAP of 50.5. The adjustment boosts the performance by 1.7 points compared with the baseline while the HOI Inferring Structure is adopted integrally.

We take two different HOI examples with the ⟨ human, surfboard ⟩ pair in Figure 6 to simply visualize the effects of three IRBs, where the brightness of pixel indicates how much the feature been noticed. Intuitively, three IRBs learn different interactive semantics from pairwise reasoning. The H-S IRB and O-S IRB exploit interactive semantics between scene and instances. The H-O IRB, 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 IRB, empirical tests have been conducted to study the effect of node numbers in interactive spaces. As summarized in Table IV, we get the best result when N is set as the value of 64.

Meanwhile, we study the impact of adopting IBAM. We compare results of 1) the baseline method, 2) concatenating F_{H-O} directly into primary branch 3) concatenating F_{H-O} to human body-part features without IBAM 4) embedding F_{H-O} into fine-grained branch by IBAM, so that we can verify the effectiveness of the IBAM. As shown in table V, concatenating F_{H-O} into primary branch brings relatively small improvement of 0.7 point compared with baseline. Whereas, embedding F_{H-O} into fine-grained branch

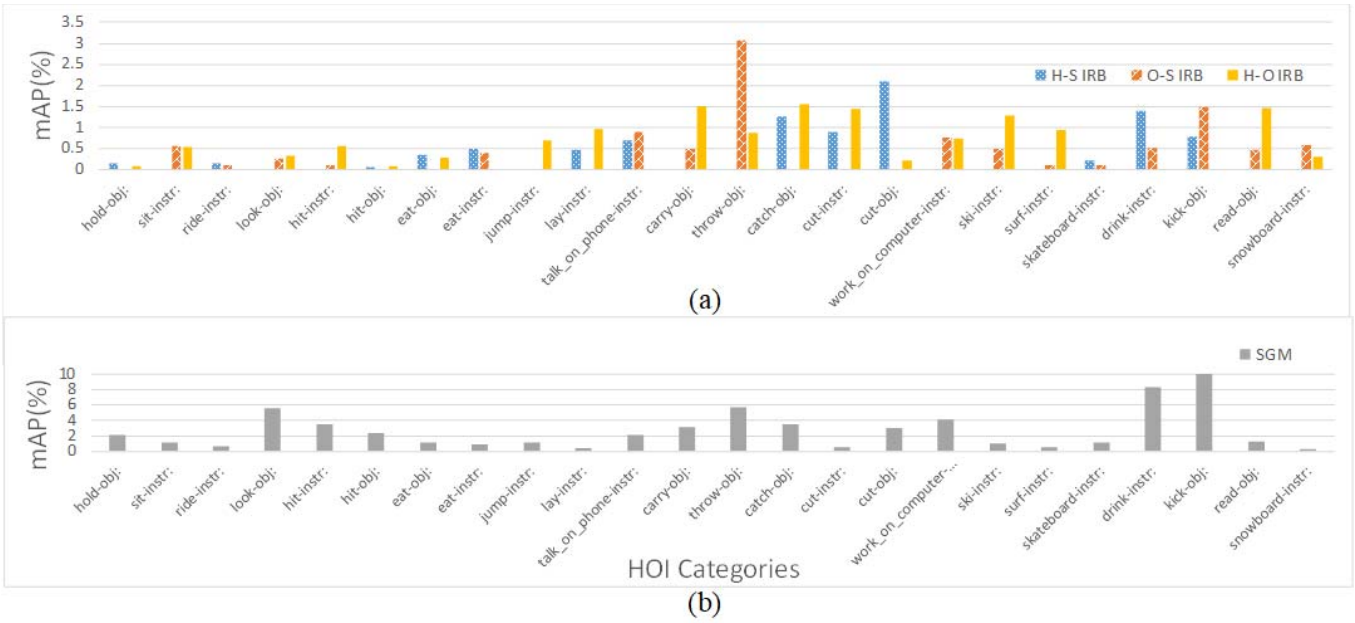

Fig. 7. (a) Class-wise analyses of Interactive Reasoning Blocks (IRBs). We show the relative contributions of three IRBs for better observation. For each HOI category, we set the minimum contribution value from three IRBs as the base value and normalize it to 0, and the other two relative contribution values are computed by subtracting the minimum contribution value. (b) Class-wise analyses of Spatial Guidance Model (SGM). We show the absolute improvements of SGM compared with baseline.

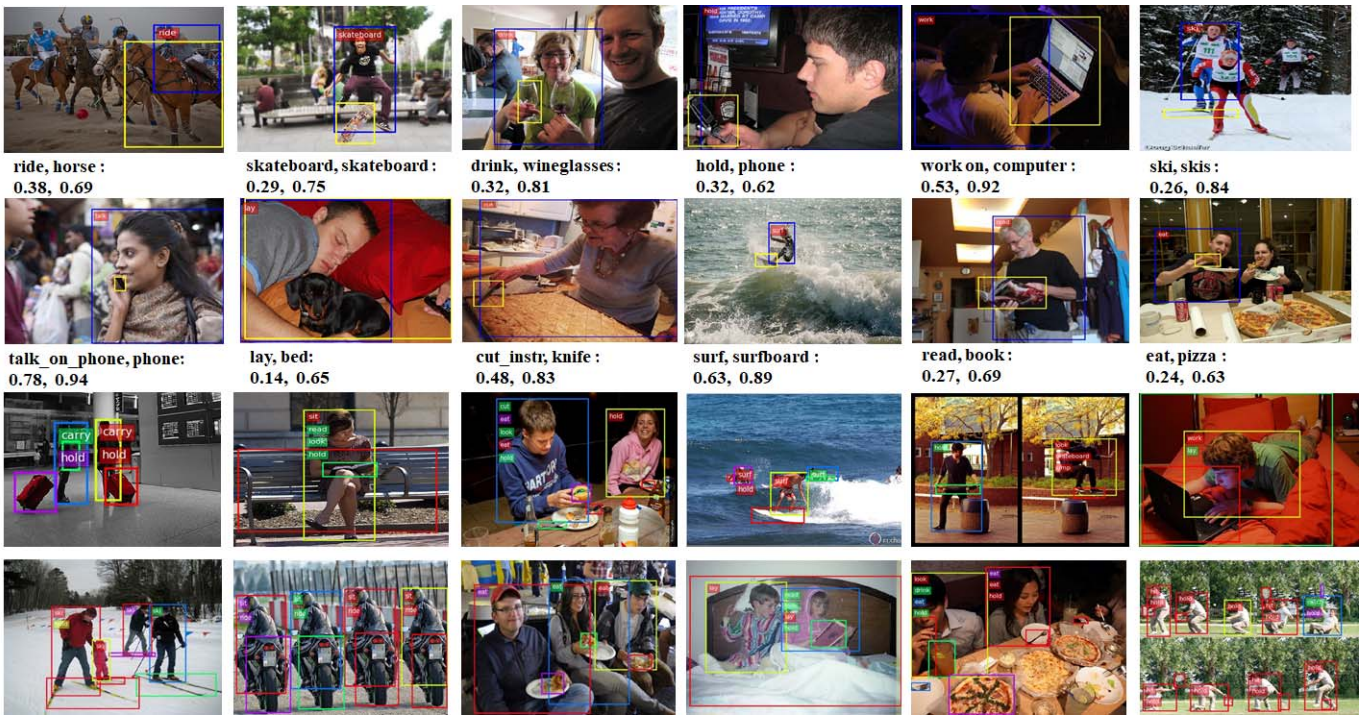

Fig. 8. Visualization of HOI detections. The first and second rows show our results compared with baseline. Texts below indicate the ⟨ verb, object ⟩ tuple and two numbers in turn represent scores predicted by baseline and our method. The third and fourth rows show multiple people take interactions with various objects concurrently detected by our method.

by IBAM brings performance gains of 1.4 points. This result shows that F_{H-O} mostly focuses on the action between human-object pair and is more effective while be embedded in fine-grained branch. On the other hand, we can see that concatenating F_{H-O} with human body-part features directly brings very limited improvement, which means F_{H-O} and human body-part feature F_p^{sgm} have similar and redundant effects. However, embedding F_{H-O} into fine-grained branch by IBAM can better improve the performance, which obtains mAP of 50.6. Therefore, as different body-parts assume

Fig. 9. Examples of incorrect detections. The subplots show examples of incorrect label (first two), hallucination of object (third) and miss-localization of object (fourth).

TABLE VI
IMPACT OF ADOPTING SPATIAL GUIDANCE MODEL

Method	$\text{mAP}_{role}(\%)$
baseline	49.2
baseline+SGM	51.1
baseline+H-O IRB with IBAM	50.6
baseline+H-O IRB with IBAM+SGM	51.5

different importance for recognizing an interaction, IBAM is effective to transmit H-O interactive semantics into human body-parts and learn the importance of each body-part to the interaction.

Moreover, we investigate the effect of adopting SGM in Table VI. Compared with the baseline, adopting SGM brings gains of 1.9 mAP. While applying the H-O IRB with IBAM to human body-part features, SGM can still contribute 0.9 mAP.

Results above indicate that each proposed component indeed contribute to final performance. Summarizing all above results validates the effectiveness of our Interactive-Net.

2) Class-Wise Analyses of Proposed Components: We further perform class-wise analyses on V-COCO to study the specific contribution of components. There are 24 HOI categories have been evaluated, in which 'cut', 'eat', 'hit' involve two types of target objects (i.e., instrument and direct object). We first show the relative contributions of three IRBs in Figure 7 (a), where the relative contribution value $AP_j^{relative}$ for each IRB can be normalized as follow:

$$AP_j^{relative} = AP_j - min(AP_{H-S}, AP_{O-S}, AP_{H-O}). \quad (12)$$

Here, AP is the absolute average precision for each HOI category. $j \in \{H-S, O-S, H-O\}$ is the index of three IRBs. We can see from Figure 7 (a) that three IRBs play different roles in class-wise detection. Specifically, we can draw a relatively conclusion from detailed class-wise results that scene-wide reasoning (i.e., H-S IRB and O-S IRB) are adept in improving detections closely related to the environments. For example, ⟨ person, throw, frisbee ⟩ (on the grass), ⟨ person, cut, cake ⟩ (in the kitchen). While instance-wide reasoning (i.e., H-O IRB) performs better in depicting interactions closely related to correlative instances (e.g., ⟨ person, carry, bag ⟩, ⟨ person, read, book ⟩).

Meanwhile, Figure 7 (b) shows the absolute improvements of SGM compared with baseline. It can be seen that SGM improves AP for all 24 HOI categories, particularly for HOIs that closely related to geometric distribution of human body-parts and object. For example, the HOI of "kick-obj" is closely associated with human legs and foot, while the HOI of "drink-instr" is highly linked to human heads and hands. Besides, some objects (e.g., soccer balls, cups) from these HOIs are too small to provide enough appearance features, but SGM can still improve the performance by utilizing the geometric distribution of human body-parts and object to enhance the visual feature learning.

D. Qualitative Examples

Several examples of detection are given in Figure 8. We first compare our results with baseline to demonstrate our improvements. Each subplot displays one detected ⟨ human, verb, object ⟩ triplet for easy observation, including the location of person and object, as well as the interaction between the above two instances. We can see from the first two rows that our method is capable of detecting various HOIs with higher scores. In addition, the third and fourth rows show that our method is able to detect multiple people taking different interactions with diversified objects. It is worth mentioning that, unlike object detection tasks that one candidate has only one ground-truth label, HOI detection tasks may contain a ⟨ human, object ⟩ pair with different labels in different situations. According to the results from Figure 8, our proposed method can adapt to complex environments to provide high quality HOI detections.

Some examples of incorrect detections are visualized in Figure 9. One of the common errors is incorrect recognition of knotty interactions. As shown in the first two subplots of Figure 9, the algorithm is more likely to predict the HOI label "hold" when it observes a luggage bag even though the interaction of "hold" does not actually happen. This phenomenon partly illustrates that modeling the finer geometric distribution of ⟨ human, object ⟩ pair is still an interesting open problem for future research. Another dominant error is related to falsely recognizing objects or humans, both hallucination and miss-localization. The object in the third subplot of Figure 9 is incorrectly detected as a computer, leading to the algorithm predicts a wrong HOI label of "work". The fourth subplot

shows an example of miss-localization of object, which is treated as a case of error in the final evaluation. These two errors could be potentially reduced by adopting better backbone networks and object detectors.

V. Conclusion

In this paper, a framework named Interactive-Net is constructed for HOI detection. Interactive-Net mainly addresses two problems in the existing methods, which are 1) lacking interactive reasoning among visual targets and 2) inefficient exploitation of geometric encoding between the human and object. Specifically, we first propose a semantic-based reasoning model named Interactive Reasoning Block (IRB), which efficiently reasons interactive semantics among visual targets by three procedures, i.e., a project function, a message passing process and an update function. Then we design a HOI Inferring Structure to integrate IRBs and parse interactive semantics in scene-wide level and instance-wide level. Moreover, a Spatial Guidance Model (SGM) based on the location of human body-parts and object are proposed, serving as the geometric guidance to promote visual feature learning. Extensive experiments and ablation studies have been conducted to evaluate our method on two public benchmarks, including V-COCO and HICO-DET. Our method outperforms existing methods with sizable gains, validating its efficacy in detecting HOIs. In future studies, we will further explore the interactive semantic reasoning and geometric encoding for HOI detection, especially improve their efficiency in modelling knotty HOIs.

Acknowledgment

The authors would like to acknowledge the AOTO-PKUSZ Joint Research Center of Artificial Intelligence on Scene Cognition Technology Innovation for its support.

References

[1] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.

[2] K. He, G. Gkioxari, P. Dollar, and R. Girshick, "Mask R-CNN," in *Proc. ICCV*, 2017, pp. 2961–2969.

[3] M. Yang, K. Yu, C. Zhang, Z. Li, and K. Yang, "DenseASPP for semantic segmentation in street scenes," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3684–3692.

[4] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 2881–2890.

[5] H.-S. Fang, S. Xie, Y.-W. Tai, and C. Lu, "RMPE: Regional multi-person pose estimation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2334–2343.

[6] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2D pose estimation using part affinity fields," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 7291–7299.

[7] S. Sharma, R. Kiros, and R. Salakhutdinov, "Action recognition using visual attention," in *Proc. Int. Conf. Learn. Represent.*, 2015, pp. 1–11.

[8] C. Feichtenhofer, A. Pinz, and A. Zisserman, "Convolutional two-stream network fusion for video action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1933–1941.

[9] S. Song, C. Lan, J. Xing, W. Zeng, and J. Liu, "An end-to-end spatio-temporal attention model for human action recognition from skeleton data," in *Proc. Nat. Conf. Artif. Intell.*, 2017, pp. 4263–4270.

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

[11] C. Gao, Y. Zou, and J. Huang, "iCAN: Instance-centric attention network for human-object interaction detection," in *Proc. Brit. Mach. Vis. Conf.*, 2018, p. 41.

[12] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

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

[14] B. Xu, J. Li, Y. Wong, Q. Zhao, and M. S. Kankanhalli, "Interact as you intend: Intention-driven human-object interaction detection," *IEEE Trans. Multimedia*, vol. 22, no. 6, pp. 1423–1432, Jun. 2020.

[15] A. Kolesnikov, A. Kuznetsova, C. Lampert, and V. Ferrari, "Detecting visual relationships using box attention," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Oct. 2019, pp. 1–5.

[16] Y.-L. Li *et al.*, "Transferable interactiveness knowledge for human-object interaction detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3585–3594.

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

[18] 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/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9469–9478.

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

[20] M. Yatskar, L. Zettlemoyer, and A. Farhadi, "Situation recognition: Visual semantic role labeling for image understanding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 5534–5542.

[21] Y. Chao, Y. Liu, X. Liu, H. Zeng, and J. Deng, "Learning to detect human-object interactions," in *Proc. Workshop Appl. Comput. Vis.*, 2018, pp. 381–389.

[22] D. Yang and Y. Zou, "A graph-based interactive reasoning for human-object interaction detection," in *Proc. 29th Int. Joint Conf. Artif. Intell.*, Jul. 2020, pp. 1111–1117.

[23] H. Idrees *et al.*, "The THUMOS challenge on action recognition for videos," *Comput. Vis. Image Understand.*, vol. 155, pp. 1–23, Feb. 2017.

[24] K. Soomro, A. Roshan Zamir, and M. Shah, "UCF101: A dataset of 101 human actions classes from videos in the wild," 2012, *arXiv:1212.0402*.

[25] G. Gkioxari, R. B. Girshick, and J. Malik, "Contextual action recognition with R*CNN," in *Proc. Int. Conf. Comput. Vis.*, 2015, pp. 1080–1088.

[26] O. Ulutan, A. S. M. Iftekhar, and B. S. Manjunath, "VSGNet: Spatial attention network for detecting human object interactions using graph convolutions," 2020, *arXiv:2003.05541*.

[27] 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. (ECCV)*, 2018, pp. 51–67.

[28] Y. Chen, Z. Wang, Y. Peng, Z. Zhang, G. Yu, and J. Sun, "Cascaded pyramid network for multi-person pose estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7103–7112.

[29] R. Girdhar and D. Ramanan, "Attentional pooling for action recognition," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 34–45.

[30] S. Jetley, N. A. Lord, N. Lee, and P. H. S. Torr, "Learn to pay attention," 2018, *arXiv:1804.02391*.

[31] D. Yang, Y. Zou, J. Zhang, and G. Li, "GID-net: Detecting human-object interaction with global and instance dependency," *Neurocomputing*, vol. 444, pp. 366–377, Jul. 2021.

[32] 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.*, Jul. 2019, pp. 5694–5702.

[33] S. Qi, W. Wang, B. Jia, J. Shen, and S. Zhu, "Learning human-object interactions by graph parsing neural networks," in *Proc. Eur. Conf. Comput. Vis.*, pp. 407–423, 2018.

[34] J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals, and G. E. Dahl, "Neural message passing for quantum chemistry," in *Proc. 34th Int. Conf. Mach. Learn.*, 2017, pp. 1263–1272.

[35] K. Cho, B. van Merrienboer, D. Bahdanau, and Y. Bengio, "On the properties of neural machine translation: Encoder-decoder approaches," 2014, *arXiv:1409.1259*.

[36] Q. Li, Z. Han, and X.-M. Wu, "Deeper insights into graph convolutional networks for semi-supervised learning," in *Proc. 32nd AAAI Conf. Artif. Intell.*, 2018, pp. 3538–3545.

[37] T.-Y. Lin *et al.*, "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.* Berlin, Germany: Springer, 2014, pp. 740–755.
[38] 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 Conf. Appl. Comput. Vis. (WACV)*, Dec. 2018, pp. 1568–1576.

Dongming Yang received the B.E. degree from Shanxi University in 2015, and the M.Sc. degree from the University of Chinese Academy of Sciences in 2018. He is currently pursuing the Ph.D. degree in computer science and engineering with Peking University. He is also an Intern with the Peng Cheng Laboratory. His research interests include computer vision and pattern recognition.

Yuexian Zou (Senior Member, IEEE) received the B.Sc. degree from the University of Electronic Science and Technology in 1985 and the Ph.D. degree from The University of Hong Kong in 2001. She is currently a Full Professor with Peking University and the Director of the Advanced Data and Signal Processing Laboratory, Peking University Shenzhen Graduate School. She was a recipient of the award Leading Figure for Science and Technology by Shenzhen Municipal Government in 2009, and also a Researcher with the Peng Cheng Laboratory. She also serves as the Deputy Director of Shenzhen Association of Artificial Intelligence (SAAI). Since 2010, she has been actively involved in teaching and research on machine learning and its applications in video and audio analysis. She conducted more than 20 research projects, including NSFC and 863 projects. She has published more than 180 academic papers in famous journals and flagship conferences, issued five invention patents, and two of them have been transferred to a company. She conducts several courses for graduate students, such as machine learning and pattern recognition, digital signal processing, and array signal processing. Her research interests mainly in machine learning for signal processing and scene understanding.

Zhu Li (Senior Member, IEEE) received the Ph.D. degree in electrical and computer engineering from Northwestern University, Evanston, in 2004. He is currently an Associate Professor with the Department of Computer Science and Electrical Engineering (CSEE), University of Missouri–Kansas City, and the Director of the NSF I/UCRC Center for Big Learning (CBL), UMKC. Prior to that, he was a AFOSR SFFP Summer Visiting Faculty with the U.S. Air Force Academy (USAFA), from 2016 to 2018 and in 2020, and with the UAV Research Center. He was a Senior Staff Researcher/Senior Manager with Samsung Research America's Multimedia Standards Research Lab, Richardson, TX, USA, from 2012 to 2015; a Senior Staff Researcher/Media Analytics Lead with FutureWei (Huawei) Technology's Media Lab, Bridgewater, NJ, USA, from 2010 to 2012; an Assistant Professor with the Department of Computing, The Hong Kong Polytechnic University, from 2008 to 2010; and a Principal Staff Research Engineer with the Multimedia Research Lab (MRL), Motorola Labs, from 2000 to 2008. His research interests include point cloud and light field compression, graph signal processing and deep learning in the next gen visual compression, image processing, and understanding. He has 47 issued or pending patents, more than 100 publications in book chapters, journals, and conferences in these areas. He serves as a Steering Committee Member of IEEE ICME (2015–2018) and an Elected Member of the IEEE Multimedia Signal Processing (MMSP), IEEE Image, Video, and Multidimensional Signal Processing (IVMSP), and IEEE Visual Signal Processing and Communication (VSPC) Tech Committees. He received the Best Paper Award at IEEE International Conference on Multimedia & Expo (ICME), Toronto, in 2006, and the Best Paper Award (DoCoMo Labs Innovative Paper) at IEEE International Conference on Image Processing (ICIP), San Antonio, in 2007. He is the Program Co-Chair for 2019 IEEE International Conference on Multimedia & Expo (ICME), and co-chaired the 2017 IEEE Visual Communication and Image Processing (VCIP). He is the Associate Editor-in-Chief of IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY and an Associate Editor of IEEE TRANSACTIONS ON IMAGE PROCESSING (since 2020), IEEE TRANSACTIONS ON MULTIMEDIA (2015–2018), and IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY (2016–2019).

Ge Li (Senior Member, IEEE) received the Ph.D. degree from the Department of Electrical Engineering, Auburn University, AL, USA, in 1999. He was a Postdoctoral Fellow with the Department of Electrical and Computer Engineering, University of California at Davis, from 2003 to 2004. After several years of research work in industry, he joined the School of Electronic and Computer Engineering, Shenzhen Graduate School, Peking University, China, in 2014, as a Full Professor. His research interests include video coding and processing, video feature extraction and analysis, and communication and signal processing.