Computational Visual Media
https://doi.org/10.1007/s41095-020-0188-2

Research Article

Detecting human–object interaction with multi-level pairwise feature network

Hanchao Liu[1], **Tai-Jiang Mu**[1] (✉), **and Xiaolei Huang**[2]

Abstract Human–object interaction (HOI) detection is crucial for human-centric image understanding which aims to infer ⟨`human, action, object`⟩ triplets within an image. Recent studies often exploit visual features and the spatial configuration of a human–object pair in order to learn the action linking the human and object in the pair. We argue that such a paradigm of pairwise feature extraction and action inference can be applied not only at the whole human and object instance level, but also at the part level at which a body part interacts with an object, and at the semantic level by considering the semantic label of an object along with human appearance and human–object spatial configuration, to infer the action. We thus propose a multi-level *pairwise feature network* (PFNet) for detecting human–object interactions. The network consists of three parallel streams to characterize HOI utilizing pairwise features at the above three levels; the three streams are finally fused to give the action prediction. Extensive experiments show that our proposed PFNet outperforms other state-of-the-art methods on the V-COCO dataset and achieves comparable results to the state-of-the-art on the HICO-DET dataset.

Keywords human–object interaction detection; pairwise feature network; deep learning; multi-level; object instance

1 Introduction

Recently, deep learning has witnessed great progress in visual recognition [1] and object detection [2–4]. To achieve deeper levels of image understanding, researchers have turned to detecting visual relationships rather than isolated instances [5, 6], a task which remains challenging due to the wide variety of relations. More specifically, detecting human-centric relationships with surrounding objects, referred to as human–object interaction (HOI) detection [7, 8], has become crucial for tasks like video understanding [9] and visual question answering [10]. The goal is to determine the ⟨`human, action, object`⟩ triplets in a single image.

Current attempts to address the problem of HOI detection usually rely on considering all ⟨`human, object`⟩ pairs in an image, where the *pairwise* features comprise three components: visual features of the human, visual features of the object, and spatial configuration linking the human and object [7, 11]. These components help to recognize actions with a typical spatial interaction pattern, e.g., `ride`, or actions strongly correlated with the presence of a person or specific objects. However, most existing methods only extract such pairwise features at the global *instance* level [12–14], which we argue is insufficient to distinguish some fine-grained actions that require subtle local cues from body parts and/or knowledge about object semantic labels—for instance, the action of eating something, involving multiple nearby objects.

This paper seeks to apply more informative pairwise representations for HOI detection in addition to the global instance information. We observe that an inherent hierarchical structure exists in pairwise features, as can be seen in Fig. 1. Beyond the *instance-level* interactions between a person and an object, there are actions strongly associated with a body part, e.g., the `hold` action involves a hand,

1 Key Laboratory of Pervasive Computing, Ministry of Education, BNRist, Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China. E-mail: H. Liu, liuhc17@mails.tsinghua.edu.cn; T.-J. Mu, taijiang@tsinghua.edu.cn (✉).

2 College of Information Sciences and Technology, Pennsylvania State University, University Park, PA 16802, USA. E-mail: suh972@psu.edu.

Manuscript received: 2020-06-26; accepted: 2020-07-20

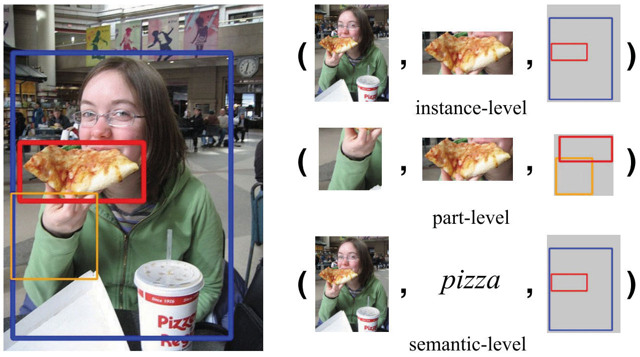


Fig. 1 HOI can be characterized using three levels of pairwise features, the instance-, body part-, and semantic-levels. At the part-level, the visually and spatially related hand and object pair indicates `hold`; at the semantic-level, the object label `pizza` strongly suggests `eat`.

and the `kick` action, a foot. Therefore, additional pairwise features at the body *part-level* that capture interactions between body parts and nearby objects can provide useful local cues for recognizing such fine-grained actions. Compared to previous part-based approaches [15, 16], our proposed part-level pairwise features are more comprehensive, consisting of three components (visual features of the body part, visual features of the object, and their relative spatial configuration), whereas previous methods do not consider all three components and are thus less capable of modeling subtle interactions between body parts and objects. Furthermore, we observe that the *semantic* label of an object can serve as a reliable prior as well as a substitute for object appearance when the object is partially occluded. Given the object semantic label, the number of visual phrases (i.e., valid pairs of action and object) becomes far smaller than the total number of ⟨`action, object`⟩ combinations. Therefore we propose a third level of pairwise features at the *semantic-level*, which utilize object labels to allow the learning of sparse correspondences between actions and objects.

In order to effectively utilize the multi-level pairwise features presented above to detect human–object interactions, we propose a novel multi-level *pairwise feature network* (PFNet) consisting of three parallel streams. PFNet aggregates pairwise visual and spatial features at three levels and incorporates both local body parts and semantic priors to achieve more robust and accurate HOI detection. The instance-level stream of PFNet captures visual and spatial configuration features of ⟨`human, object`⟩ pairs. The part-level stream captures visual and spatial relationship features of ⟨`body-part, object`⟩ pairs. Specifically, at the part-level, we enlarge the receptive field of the object visual feature to be the union of the bounding boxes covering the object and a neighbouring body part. The part-level spatial configuration is represented by the distance between the object and its nearest body part. The semantic-level stream resembles the instance-level counterpart but captures pairwise relations by replacing the object visual feature by its semantic label feature. Lastly, the three streams are fused to predict the HOI. A comparison with other methods conducted on two large-scale datasets, V-COCO [17] and HICO-DET [7], shows that our method achieves state-of-the-art performance on V-COCO and the best result on HICO-DET, without needing any extra annotation.

2 Related work

Action recognition is a human-centric visual recognition task closely related to HOI detection. Action recognition usually relies on pose-guided human appearance and contextual information. Zhao et al. [18] generate body parts with the assistance of a human pose estimation network and use the state of body parts to complement global human appearance. Luvizon et al. [19] conduct multi-task learning for both action recognition and pose estimation to improve the performance for each task. Attention mechanisms are also widely used for action recognition. Abdulmunem et al. [20] extract both local and global descriptors for efficient action recognition guided by object saliency. Girdhar and Ramanan [21] propose a top–down and bottom–up attention mechanism to capture global context and local features. Although action recognition can be considered as an image-level task, these strategies can readily be transferred to recognizing and detecting instance-level human–object interactions. In our work, we also utilize detailed information about body parts to enrich global features.

Human–object interaction detection lies at the intersection of action recognition and general visual relationship detection. In existing instance-level approaches, a multi-stream network extracts pairwise visual and spatial features for a human–object pair for interaction prediction. In addition, instance-centric attention [11], or spatial relation guided attention

[22], can be used to refine the pairwise features. Wang et al. [13] introduce context-aware human and object appearance features that better incorporate information from background scenes. Some networks further predict a binary interaction score [12, 15] for a human–object pair or estimate the object location with a localization branch [8]. HOIs can also be parsed as a scene graph [5] so that information from all human–object pairs in one image can be utilized. Instance features are refined using iterative message passing [23] or graph convolution [22].

PMFNet [15] and RPNN [16] are two typical part-based approaches that utilize visual features of body parts. PMFNet has a zoom-in module that extracts local visual and spatial features from pose keypoint guided regions. RPNN has a graph for human and body parts and another graph for object and body parts, which enrich the coarse instance-level human and object features with weighted body part visual features. However, they capture a certain scope of local feature which is suboptimal for representing interactions between body parts and objects. Unlike these approaches, we employ the same type of pairwise representation for local part-level features as for instance-level pairs.

Moreover, semantic features carried by action and object labels have also been explored to obtain better generalization when few examples exist for an HOI category [6, 24–26]. A common approach is to learn a joint embedding space that matches visual and language features of HOIs [6, 24, 25] so that a similarity term can be appended to the final prediction score to determine how an action prediction matches its semantic meanings. Instead of learning a joint embedding space, we directly model semantic dependency for action categories based on object labels.

Finally, very recently models built on anchor-free point-based detection frameworks have been proposed to perform HOI detection [27, 28]. They treat HOIs as keypoints lying between a human and an object.

3 Multi-level pairwise feature network

3.1 Network architecture

We adopt a two-stage pipeline consisting of an instance localization stage and an interaction recognition stage, following Refs. [11, 12]. Given an image I, an off-the-shelf object detector, e.g., Faster R-CNN [2] with a ResNet50 backbone, first detects all human instances with bounding boxes b_h, and all object instances with bounding boxes b_o and class labels c_o. The feature map F for the image I is extracted from the ResNet50 C4 *conv* layer. Meanwhile, a human pose estimation network parses human instances h into keypoints k_h for extracting body part boxes b_h^k. We perform action prediction on all pairs of humans and objects.

We regard a human–object interaction as a function of pairwise features at multiple levels: instance level, body part level, and semantic level. As shown in Fig. 2, the network extracts the instance-level pairwise feature f_{ins} by pairing b_h, b_o to capture interactions on a global scale. The part-level pairwise feature f_{par} is extracted from b_h^k and b_o to provide finer-grained details about interactions between local body parts and objects. The network also extracts a semantic-level pairwise feature f_{sem} from b_h, b_o, and the object label c_o which carries prior semantic information. The pose skeleton k_h is also used when constructing f_{ins} and f_{sem}. The final interaction score is obtained using a factorization form. Specifically, we concatenate f_{ins} and f_{par} to predict an action probability $p_{h,o}^a$ based on appearance, while at the same time we use f_{sem} to predict an action probability $q_{h,o}^a$ with semantic prior. The final HOI score $s_{h,o}^a$ relating a human h and an object o with action a is the product of the two terms:

$$s_{h,o}^a = p_{h,o}^a q_{h,o}^a \tag{1}$$

Next we detail how we extract the three levels of pairwise features through three parallel streams. Each stream follows a similar concise pattern with minor structural changes to adapt it to the variations between each level.

3.2 Instance-level pairwise feature stream

The instance-level pairwise feature captures the holistic visual and spatial relationships for a human and object pair [7, 11–13, 15]. To obtain this pairwise feature, we first crop visual features of the human instance and the object instance in the pair by applying the RoI-Align [29] operation on the feature map F. Then we apply global average pooling (GAP), followed by two fully connected layers, to obtain the feature vectors f_{ins}^h and f_{ins}^o. The spatial configuration can be represented as a three-channel spatial-pose map [12, 15] which augments the two-channel binary mask map of human and objects [7, 11] with an

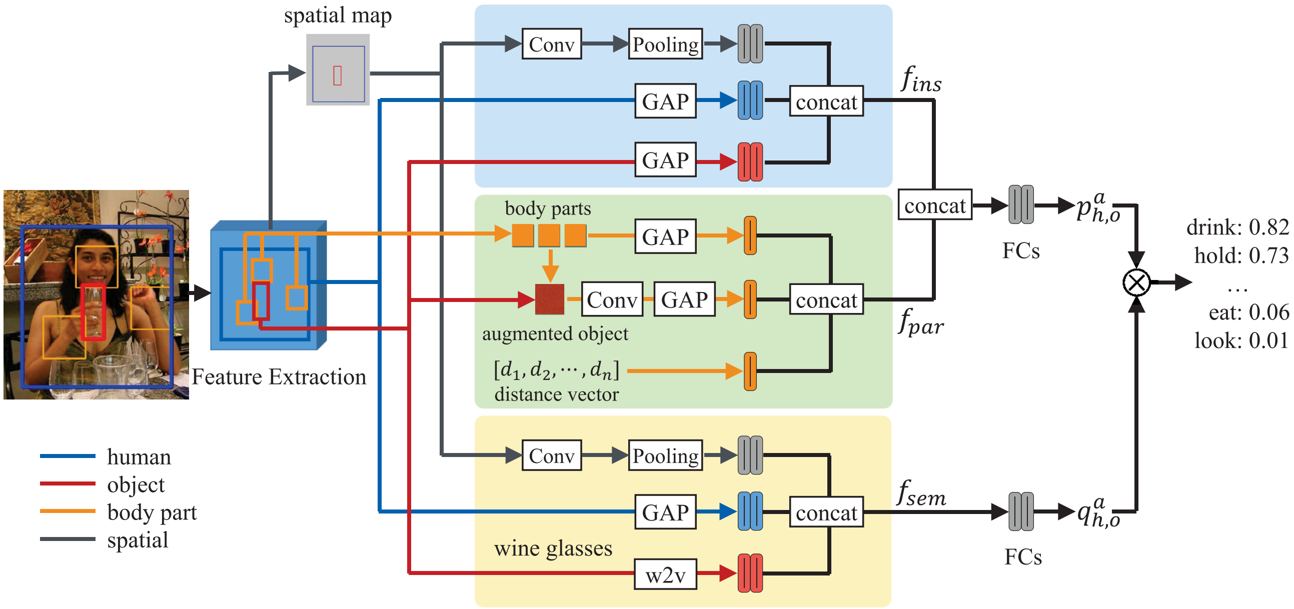


Fig. 2 Architecture of our pairwise feature network. Human boxes, object boxes, and body part boxes extracted from the input image are fed to three streams to learn instance-level, part-level, and semantic-level pairwise features, respectively, using both visual and spatial information. The final action prediction is obtained by score fusion.

additional coarse pose skeleton layout. Specifically, we generate a tight bounding box $b_{h,o}$ enclosing the human and object bounding boxes b_h and b_o. We fill the region of $b_{h,o}$ with the human and object masks in the first and second channel, respectively. In the third channel we draw pose keypoints in k_h as the body joints and lines linking them as the body skeleton. The lines have gray values ranging from 0.15 to 0.95 in order to encode different parts. The human pose is obtained as in Ref. [30] and the body skeleton follows a conventional pattern [12]. After resizing the spatial-pose map to a size of $64 \times 64 \times 3$, we use two *conv* layers followed by max pooling layers and two fully connected layers to obtain the spatial feature vector $f^{\text{sp}}_{\text{ins}}$.

Considering both visual and spatial relations, the instance-level pairwise feature f_{ins} can be represented as a concatenation of the feature vector f^h_{ins}, f^o_{ins}, and $f^{\text{sp}}_{\text{ins}}$:

$$f_{\text{ins}} = f^h_{\text{ins}} \oplus f^o_{\text{ins}} \oplus f^{\text{sp}}_{\text{ins}} \qquad (2)$$

where $\oplus$ denotes concatenation of feature vectors.

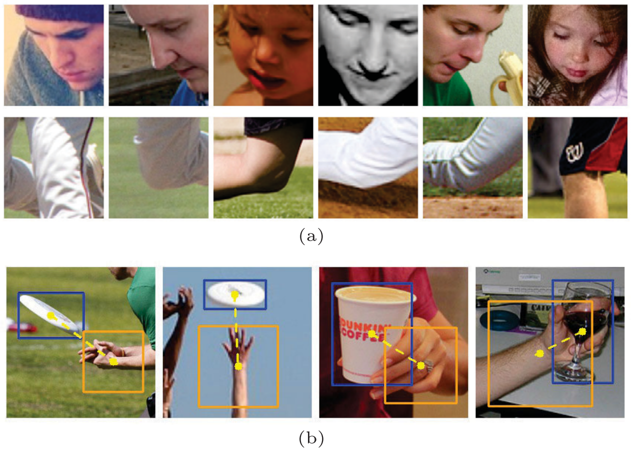
3.3 Part-level pairwise feature stream

The part-level pairwise feature stream is responsible for capturing local interactions between objects and body parts. To this end, we first consider visual feature and spatial relations for a body part and object pair. We then organize a set of part-level pairwise features into a form like Eq. (2) with aggregated body part feature, augmented object feature and aggregated spatial configuration.

Given a set of human pose keypoints k_h, we crop $n = 10$ body part regions whose center points can be well defined by k_h, following Ref. [31]. Specifically, the ten body parts include head, pelvis, both left and right arms, both hands, both knees, and both feet. All body-part regions are square boxes of size proportional to the height of the human bounding box (Fig. 3(a)). Here we denote a body part region as $b^{k_i}_h$, where $k_i \in k_h$. We apply RoI-Align to crop a body part feature from the feature map according to $b^{k_i}_h$, followed by global average pooling to generate a feature $f^h_{k_i}$ for each body part. Body part visual features are aggregated using a fully connected layer W^h_{par} to get part-level human visual feature f^h_{par}:

$$f^h_{\text{par}} = W^h_{\text{par}}(f^h_{k_1} \oplus f^h_{k_2} \oplus \cdots \oplus f^h_{k_n}) \qquad (3)$$

The original object bounding box often has a limited receptive field to capture important visual cues about how an object interacts with a local body part. However, this is crucial for recognizing actions when a body part and an object have direct contact or are close to each other. To address this issue we introduce an augmented object feature that also includes a neighbouring body part. Given the fact that a number of actions involve local interaction between hands and objects, we consider a hand-


(a)

(b)

Fig. 3 Examples of part-level pairwise features. (a) Body part regions, e.g., head state (above) for `read` and knee state (below) for `throw`. (b) We augment the object feature by cropping the feature from the expanded union box covering the object (blue) and its neighbouring body part (e.g., hand) to enrich the visual cues for local interactions, e.g., `throw` and `hold`, which are strongly correlated with body parts. The distance between the object and body part (dashed lines) is utilized as a local spatial feature.

augmented object feature. Specifically, as shown in Fig. 3(b), we first find the closest hand part box to an object (for left or right hand). Then we generate a union box covering that hand part and the object, and expand the region by a margin. Similarly, we adopt RoI-Align to crop the feature which is followed by a *conv* layer, a GAP, and a fully connected layer W^{o}_{par} to extract the hand augmented object feature f^{o}_{par}. Note that, although we augment the object feature with a specific type of body part here, one could also consider an arbitrary group of body parts.

As the size of a body part bounding box does not have specific meaning, we utilize body part to object distance as a discriminant feature to indicate which body part has a close spatial relation with the object. We use the normalized distance between the object center (u_o, v_o) and a body part box center (u_i, v_i):

$$d_i = D\left((u_i/W, v_i/H), (u_o/W, v_o/H)\right) \tag{4}$$

where H, W are the height and width of the union bounding box of human b_h and object b_o, and $D(\cdot,\cdot)$ is the Euclidean distance between two two-dimensional (2D) points. Considering all distances, we have a distance vector and then obtain the spatial feature $f^{\text{sp}}_{\text{par}}$ by applying another fully connected layer $W^{\text{sp}}_{\text{par}}$:

$$f^{\text{sp}}_{\text{par}} = W^{\text{sp}}_{\text{par}}([d_1, \cdots, d_n]) \tag{5}$$

Finally, the part-level pairwise feature is represented by concatenating the aggregated local body part visual feature, augmented object feature and local spatial relations:

$$f_{\text{par}} = f^{h}_{\text{par}} \oplus f^{o}_{\text{par}} \oplus f^{\text{sp}}_{\text{par}} \tag{6}$$

We do not employ attention modules for feature refinement [13, 15], as the pose keypoints are already effective for region selection. The effectiveness of each part is validated in Section 4.5.2.

Since the instance-level and part-level pairwise features together encode the appearance of a human–object pair, we therefore concatenate them and pass the result through fully connected layers W_{appr} to predict an action probability $p^{a}_{h,o}$:

$$p^{a}_{h,o} = \sigma(W_{\text{appr}}(f_{\text{ins}} \oplus f_{\text{par}})) \tag{7}$$

where σ is a sigmoid layer.

3.4 Semantic-level pairwise feature stream

Inspired by Ref. [26] which utilizes a group of semantically related object labels to improve generalization, we propose a semantic-level pairwise feature incorporating object labels to explore semantic dependency for different actions.

The semantic-level pairwise feature for a human–object pair is constructed in the same way as the instance-level pairwise feature except that the visual object feature is replaced by the language embedding feature of its object label. This is based on the observation that given the human appearance, human–object spatial relations, and object labels, very reasonable predictions can be made in scenarios like `eat` or `drink`. Thus, the semantic-level pairwise feature is defined as

$$f_{\text{sem}} = f^{h}_{\text{sem}} \oplus f^{o}_{\text{sem}} \oplus f^{\text{sp}}_{\text{sem}} \tag{8}$$

We adopt a weight sharing strategy to learn this feature. Weights for the human visual feature and spatial relation are shared across instance-level and semantic-level features for joint learning, for better consistency. Therefore we have f^{h}_{sem} sharing parameters with f^{h}_{ins} and $f^{\text{sp}}_{\text{sem}}$ sharing parameters with $f^{\text{sp}}_{\text{ins}}$; f^{h}_{ins} and $f^{\text{sp}}_{\text{ins}}$ were described in Section 3.2. The semantic feature f^{o}_{sem} is obtained from object labels c_o with three fully connected layers, the first of which is initialized by word2vec [32] embeddings.

We utilize f_{sem} to independently predict an action classification score $q^{a}_{h,o}$ as a semantic prior:

$$q^{a}_{h,o} = \sigma(W_{\text{sem}}(f_{\text{sem}})) \tag{9}$$

where σ is a sigmoid layer and W_{sem} are two fully connected layers.

3.5 Loss function

Our proposed network can be trained in an end-to-end fashion. In an HOI detection task, a person can simultaneously conduct more than one action, making it a multi-label classification problem. As positive ⟨`human, action, object`⟩ triplets are relatively sparse among all triplets, some previous work [12, 15] further predict an interaction or affinity term to filter out human–object pairs that are not interacting. Here we address the sparsity problem by applying a focal loss [33] which adaptively changes the weights of easy negative samples and hard positive samples. In detail, for a human h and an object o in an image, we calculate $s^a_{h,o}$ according to Eq. (1). The loss for a prediction $s^a_{h,o}$ with ground truth label $y^a_{h,o}$ is expressed as

$$\mathcal{L}(s^a_{h,o}, y^a_{h,o}) = y^a_{h,o}(1 - s^a_{h,o})^2 \log s^a_{h,o} + (1-y^a_{h,o})(s^a_{h,o})^2 \log(1 - s^a_{h,o}) \quad (10)$$

The total loss for an image is the sum of $\mathcal{L}(s^a_{h,o}, y^a_{h,o})$ over all human, object, and action dimensions.

4 Experiments

In this section we first introduce datasets and evaluation metrics used in our experiments. Then comparisons with state-of-the-art methods are presented; we also conducted extensive experiments to validate the effectiveness of our proposed network.

4.1 Datasets

V-COCO [17] and HICO-DET [7] are the two most commonly used benchmark datasets for HOI. V-COCO annotates 10,346 human instances with 26 actions for a subset of the COCO [34] dataset. It has 2533 images for training, 2867 images for validation, and 4946 images for testing. HICO-DET is a much larger dataset with 117 action labels, 80 object labels in COCO, and altogether 600 HOI classes. The training set has 38,118 images and test set has 9658 images. The whole dataset annotates more than 150,000 instances.

4.2 Evaluation metrics

We use standard mean average precision (mAP) as evaluation metric for both datasets. A ⟨`human, action, object`⟩ result is regarded as a true positive if the action is correctly predicted and the intersection over union (IoU) between the detected human/object instance and the ground truth instance is greater than 0.5.

4.3 Implementation details

To enable a fair comparison we adopt the same setting as Ref. [12]. Faster R-CNN ResNet50 [2] pre-trained on the COCO dataset is used as the feature backbone and kept frozen. The pose estimation result is obtained using AlphaPose [30] and the object detection result comes from the Detectron [35]. In each level of pairwise features, f_h, f_o, and f_{sp} are all 1024 dimensional features, giving the overall pairwise feature a size of 3072. The number of hidden units for all fully connected layers is set to 1024. In the part-level feature, the dimension of body part features is reduced to 256 with a spatial size of 5 × 5. Following Ref. [12], we train V-COCO on the *trainval* set. During training, we sample positive and negative samples with a ratio of 1:3 using 8 training images as a batch and use the Adam optimizer [36]. We set the initial learning rate to 10^{-4} and reduce it to 10^{-5} in the 11th epoch for V-COCO and in the 7th epoch for HICO-DET. Our model is trained for 20 epochs in total for both datasets. During testing, the object threshold is set to 0.1 while the human threshold is set to 0.3 for V-COCO and 0.5 for HICO-DET. As the whole pipeline starts with a localization stage, the quality of detected human and object instances affects the final HOI detection score. Therefore the final score for action prediction is merged with the instance confidence scores. We apply the Low-grade Instance Suppression (LIS) function [12] to make a non-linear adjustment to the original detection scores. For V-COCO we also conduct post-processing to remove contradictory predictions, following Ref. [12].

4.4 Comparison with state of the art

We report quantitative results from our proposed pipeline on the V-COCO dataset and compare them to the results from other state-of-the-art methods in Table 1. One can see that our proposed approach achieves a $\mathrm{mAP}_{\mathrm{role}}$ of 52.8, surpassing all other methods. Table 2 compares results from a number of approaches, using COCO-pretrained detectors, for the HICO-DET dataset. HICO-DET has two different settings, *Default* and *Known objects*. For each setting, the model is evaluated in three different modes—the full mode with all 600 HOIs, the rare mode with 138 HOIs that have fewer than 10 training

Table 1 Comparison with state-of-the-art methods on the V-COCO [17] test set

Method	Backbone	$\mathrm{mAP}_{\mathrm{role}}$
Gupta et al. [17] ([8] impl.)	ResNet50-FPN	31.8
InteractNet [8]	ResNet50-FPN	40.0
GPNN [23]	ResNet50	44.0
iCAN [11]	ResNet50	45.3
Xu et al. [24]	ResNet50	45.9
Wang et al. [13]	ResNet50	47.3
RPNN [16]	ResNet50	47.5
Li et al. [12]	ResNet50	48.7
Zhou et al. [37]	ResNet50	48.9
Wang et al. [27]	Hourglass104	51.3
VSGNet [22]	ResNet152	51.8
PMFNet [15]	ResNet50-FPN	52.0
Ours	ResNet50	**52.8**

Table 2 Comparison with state-of-the-art methods on the HICO-DET [7] test set.

Method	Default			Known object		
	Full	Rare	Non-rare	Full	Rare	Non-rare
HO-RCNN [7]	7.81	5.37	8.54	10.41	8.94	10.85
Shen et al. [38]	6.46	4.24	7.12	—	—	—
InteractNet [8]	9.94	7.16	10.77	—	—	—
GPNN [23]	13.11	9.34	14.23	—	—	—
iCAN [11]	14.84	10.45	16.15	16.26	11.33	17.73
Xu et al. [24]	14.70	13.26	15.13	—	—	—
Wang et al. [13]	16.24	11.16	17.75	17.73	12.78	19.21
Gupta et al. [14]	17.18	12.17	18.68	—	—	—
Li et al. [12]	17.22	13.51	18.32	19.38	15.38	20.57
RPNN [16]	17.35	12.78	18.71	—	—	—
PMFNet [15]	17.46	15.65	18.00	20.34	17.47	21.20
Peyre et al. [6]	19.40	15.40	20.75	—	—	—
Wang et al. [27]	19.56	12.79	**21.58**	22.05	15.77	23.92
VSGNet [22]	19.80	16.05	20.91	—	—	—
Ours	**20.05**	**16.66**	21.07	**24.01**	**21.09**	**24.89**

samples, and the non-rare mode with the remaining 462 HOIs. We report a very competitive mAP of 20.05 for Default mode. Note that we outperform PMFNet [15] which is another pose-guided multi-level network by a large margin (2.59) due to better pairwise feature representation and training strategies. While two recently published methods [39, 40] achieve 21.34 and 22.65 respectively, they rely on external 3D information and heavily annotated body part states. Figure 4 provides some sample qualitative results for V-COCO and HICO-DET datasets. As can be seen, our model distinguishes well the fine-grained actions and is able to handle challenging cases in which multiple humans interact with multiple objects.

4.5 Ablation study

4.5.1 Effect of each level of pairwise features

Our network consists of three levels of pairwise features. To fully understand how they contribute to the final result, we conducted ablation experiments using the V-COCO dataset. We used the model with only the instance-level feature as the baseline and ablated the other two pairwise features. All models were trained with the same settings. Results are shown in Table 3. The instance-level baseline model achieves a $\mathrm{mAP}_{\mathrm{role}}$ of 49.2. Adding a part-level pairwise feature improves the performance by 2.4, while adding a semantic-level pairwise feature improves the result by 1.4. Adding both together yields an absolute gain of 3.6 over the baseline, demonstrating that all levels of pairwise features benefit performance.

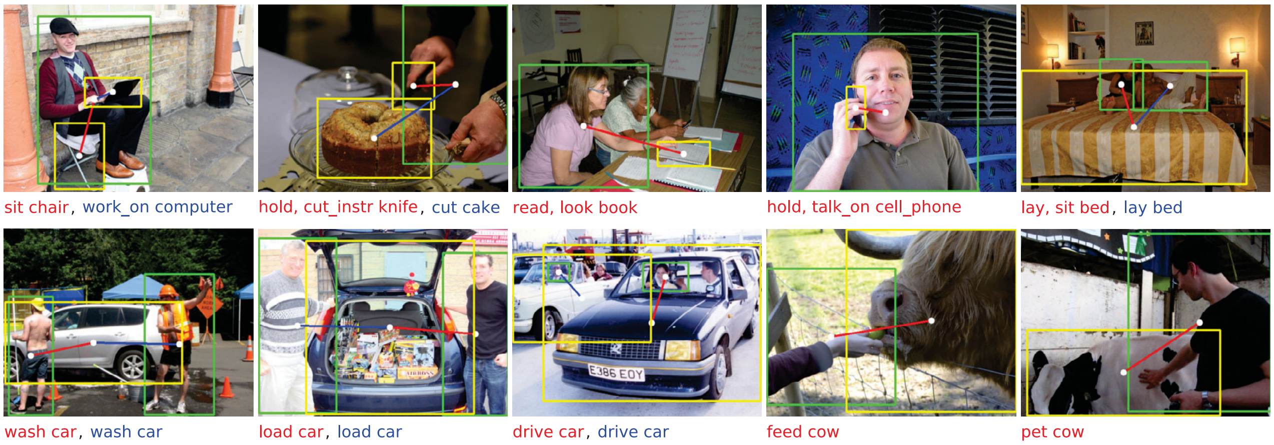

sit chair, work_on computer
hold, cut_instr knife, cut cake
read, look book
hold, talk_on cell_phone
lay, sit bed, lay bed
wash car, wash car
load car, load car
drive car, drive car
feed cow
pet cow

Fig. 4 HOI detection examples from V-COCO (above) and HICO-DET (below). Green boxes: humans. Yellow boxes: objects. Interaction is indicated by a colored line between human and object center points.

Table 3 Contribution of different levels of pairwise features to performance, for the V-COCO dataset

f_{ins}	f_{par}	f_{sem}	mAP_{role}
✓	✗	✗	49.2
✓	✓	✗	51.6
✓	✗	✓	50.6
✓	✓	✓	52.8

Table 4 Per-class performance with and without the semantic-level pairwise feature, for the V-COCO test set

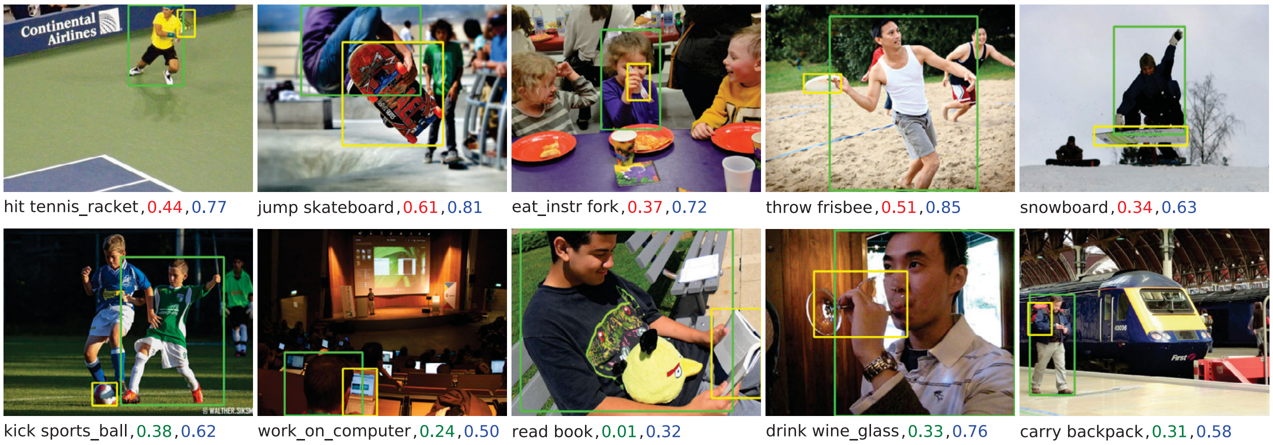
Action class	Without	With
hold-obj	42.3	42.7
sit-instr	30.1	30.0
ride-instr	72.4	71.9
look-obj	40.7	41.7
hit-instr	77.5	76.2
hit-obj	48.5	49.6
eat-obj	40.8	40.6
eat-instr	7.1	**9.0**
jump-instr	56.1	55.9
lay-instr	32.4	32.0
talk_on_phone-instr	57.2	55.6
carry-obj	42.3	**46.1**
throw-obj	46.4	46.5
catch-obj	50.0	49.9
cut-instr	45.7	46.1
cut-obj	40.7	41.1
work_on_computer-instr	67.7	**69.4**
ski-instr	51.2	51.5
surf-instr	81.1	82.0
skateboard-instr	87.7	88.2
drink-instr	38.3	**48.1**
kick-obj	71.2	**74.7**
point-instr	0.7	0.1
read-obj	35.8	**43.7**
snowboard-instr	75.1	76.0
AP (omit point)	51.6	52.8

We also investigated the necessity of predicting a pairwise interaction or affinity score. Using our full model, we applied an interaction score pretrained on HICO-DET, provided by Li et al. [12], to filter out non-interacting pairs. The final performance improved only very slightly by 0.08, indicating that our model has implicitly captured the pairwise affinity.

We also evaluated the per-class performance to examine the effect of the semantic-level pairwise feature. As shown in Table 4, the model using a pairwise semantic feature significantly outperforms one without semantics on specific action classes. Actions like `drink` and `read` can be well predicted with the assistance of a class-specific action prior. This demonstrates the efficacy of the semantic-level feature. Figure 5 shows various cases in which predictions are considerably improved by employing part-level and semantic-level features.

4.5.2 Components in part-level feature

As a pairwise feature exploits both visual and spatial information, we also investigated the contribution of each component to the part-level feature. We considered various combinations of aggregated body part feature f^h_{par}, augmented object feature f^o_{par}, and part-level spatial feature $f^{\text{sp}}_{\text{par}}$; results are shown in Table 5. The individual component features improve the result by 0.9, 0.9, and 0.8, respectively. Dropping any of the component features causes the final performance to degrade by 0.7, 0.5, and 0.4 respectively, indicating that all components are helpful.



hit tennis_racket,0.44,0.77 jump skateboard,0.61,0.81 eat_instr fork,0.37,0.72 throw frisbee,0.51,0.85 snowboard,0.34,0.63

kick sports_ball,0.38,0.62 work_on_computer,0.24,0.50 read book,0.01,0.32 drink wine_glass,0.33,0.76 carry backpack,0.31,0.58

Fig. 5 Examples from V-COCO showing effectiveness of multi-level pairwise features. Red, green, and blue scores are results from the baseline model with the instance-level feature only, with instance-level and part-level features, and all features, respectively.

Table 5 Effect of each component of the part-level pairwise feature, for the V-COCO test set

f^h_{par}	f^o_{par}	$f^{\mathrm{sp}}_{\mathrm{par}}$	$\mathrm{mAP}_{\mathrm{role}}$
✗	✗	✗	50.6
✓	✗	✗	51.5
✗	✓	✗	51.5
✗	✗	✓	51.4
✓	✓	✗	52.4
✓	✗	✓	52.3
✗	✓	✓	52.1
✓	✓	✓	52.8

5 Limitations

Our approach has limitations. Firstly, as shown in Table 4, our approach performs worse with the semantic-level pairwise feature for some action classes such as “talk on phone”. This is mainly because the semantic prior may lead to an incorrect association between human and object in confusing scenes: see Fig. 6(left). A possible solution could be to apply attention modules for level-wise feature selection to weight different features. Secondly, our approach is two-staged and the results are influenced by accuracy of object detection: see Fig. 6(right). An end-to-end multi-task network that simultaneously detects objects and interactions could help to improve both accuracy and efficiency.

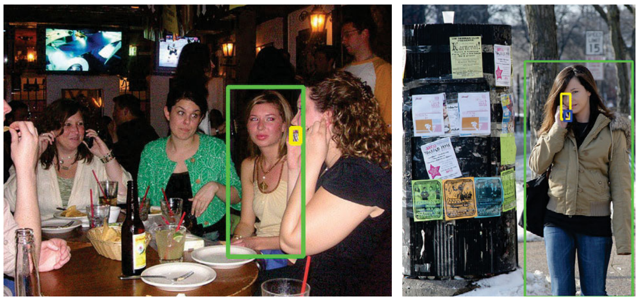


Fig. 6 Failures. Left: human and object are wrongly associated in a confusing scene. Right: the object detector incorrectly localizes the object. Yellow, blue boxes: detected, ground truth objects, respectively. Green box: detected human.

6 Conclusions

In this paper, we have presented a multi-level pairwise feature network (PFNet) for human–object interaction detection. We represent the human–object interaction as multi-level pairwise visual and spatial relations in a unified formulation. In addition to the instance-level pairwise feature, the part-level pairwise feature exploits local visual and spatial relations between a body part and an object guided by pose keypoints, while the semantic-level pairwise feature represents an object using its semantic label. Extensive experiments show that our proposed approach utilizing multi-level pairwise features for HOI detection outperforms other methods on the V-COCO dataset, while various ablation studies demonstrate the utility of multi-level pairwise features and fine-grained visual and spatial features involving body parts.

Acknowledgements

We thank the reviewers for their constructive comments. This work was supported by the National Natural Science Foundation of China (Project No. 61902210), a Research Grant of Beijing Higher Institution Engineering Research Center, and the Tsinghua–Tencent Joint Laboratory for Internet Innovation Technology.

References

[1] He, K. M.; Zhang, X. Y.; Ren, S. Q.; Sun, J.; Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 770–778, 2016.

[2] Ren, S. Q.; He, K. M.; Girshick, R.; Sun, J. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* Vol. 39, No. 6, 1137–1149, 2017.

[3] Redmon, J.; Farhadi, A. YOLO9000: Better, faster, stronger. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 6517–6525, 2017.

[4] Borji, A.; Cheng, M. M.; Hou, Q. B.; Jiang, H. Z.; Li, J. Salient object detection: A survey. *Computational Visual Media* Vol. 5, No. 2, 117–150, 2019.

[5] Xu, D. F.; Zhu, Y. K.; Choy, C. B.; Fei-Fei, L. Scene graph generation by iterative message passing. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 3097–3106, 2017.

[6] Peyre, J.; Laptev, I.; Schmid, C.; Sivic, J. Detecting unseen visual relations using analogies. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 1981–1990, 2019.

[7] Chao, Y. W.; Liu, Y. F.; Liu, X. Y.; Zeng, H. Y.; Deng, J. Learning to detect human–object interactions. *arXiv preprint* arXiv:1702.05448, 2017.

[8] Gkioxari, G.; Girshick, R.; Dollár, P.; He, K. M. Detecting and recognizing human–object interactions. *arXiv preprint* arXiv:1704.07333, 2017.

[9] Ma, C. Y.; Kadav, A.; Melvin, I.; Kira, Z.; AlRegib, G.; Graf, H. P. Attend and interact: Higher-order object interactions for video understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 6790–6800, 2018.

[10] Mallya, A.; Lazebnik, S. Learning models for actions and person–object interactions with transfer to question answering. In: *Computer Vision—ECCV 2016. Lecture Notes in Computer Science, Vol 9905*. Leibe, B.; Matas, J.; Sebe, N.; Welling, M. Eds. Springer Cham, 414–428, 2016.

[11] Gao, C.; Zou, Y. L.; Huang, J. B. iCAN: Instance-centric attention network for human–object interaction detection. *arXiv preprint* arXiv:1808.10437, 2018.

[12] Li, Y. L.; Zhou, S. Y.; Huang, X. J.; Xu, L.; Ma, Z.; Fang, H. S.; Wang, Y. F.; Lu, C. W. Transferable interactiveness knowledge for human–object interaction detection. *arXiv preprint* arXiv:1881.08264, 2019.

[13] Wang, T. C.; Anwer, R. M.; Khan, M. H.; Khan, F. S.; Pang, Y. W.; Shao, L. et al. Deep contextual attention for human–object interaction detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 5693–5701, 2019.

[14] Gupta, T.; Schwing, A. G.; Hoiem, D. No-frills human–object interaction detection: Factorization, layout encodings, and training techniques. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 9676–9684, 2019.

[15] Wan, B.; Zhou, D. S.; Liu, Y. F.; Li, R. J.; He, X. M. Pose-aware multi-level feature network for human object interaction detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 9468–9477, 2019.

[16] Zhou, P.; Chi, M. Relation parsing neural network for human–object interaction detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 843–851, 2019.

[17] Gupta, S.; Malik, J. Visual semantic role labeling. *arXiv preprint* arXiv:1505.04474, 2015.

[18] Zhao, Z. C.; Ma, H. M.; You, S. D. Single image action recognition using semantic body part actions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 3411–3419, 2017.

[19] Luvizon, D. C.; Picard, D.; Tabia, H. 2D/3D pose estimation and action recognition using multitask deep learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 5137–5146, 2018.

[20] Abdulmunem, A.; Lai, Y. K.; Sun, X. F. Saliency guided local and global descriptors for effective action recognition. *Computational Visual Media* Vol. 2, No. 1, 97–106, 2016.

[21] Girdhar, R.; Ramanan, D. Attentional pooling for action recognition. *arXiv preprint* arXiv:1711.01467, 2017.

[22] Ulutan, O.; Iftekhar, A. S. M.; Manjunath, B. S. VSGNet: Spatial attention network for detecting human object interactions using graph convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 13617–13626, 2020.

[23] Qi, S. Y.; Wang, W. G.; Jia, B. X.; Shen, J. B.; Zhu, S. C. Learning human–object interactions by graph parsing neural networks. In: *Computer Vision—ECCV 2018. Lecture Notes in Computer Science, Vol. 11213.* Ferrari, V.; Hebert, M.; Sminchisescu, C.; Weiss, Y. Eds. Springer Cham, 407–423, 2018.

[24] Xu, B.; Wong, Y.; Li, J.; Zhao, Q.; Kankanhalli, M. S. Learning to detect human–object interactions with knowledge. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019–2028, 2019.

[25] Kato, K.; Li, Y.; Gupta, A. Compositional learning for human object interaction. In: *Computer Vision—ECCV 2018. Lecture Notes in Computer Science, Vol. 11218.* Ferrari, V.; Hebert, M.; Sminchisescu, C.; Weiss, Y. Eds. Springer Cham, 247–264, 2018.

[26] Bansal, A.; Rambhatla, S. S.; Shrivastava, A.; Chellappa, R. Detecting human–object interactions via functional generalization. In: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 34, No. 7, 10460–10469, 2020.

[27] Wang, T. C.; Yang, T.; Danelljan, M.; Khan, F. S.; Zhang, X. Y.; Sun, J. Learning human–object interaction detection using interaction points. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 4115–4124, 2020.

[28] Liao, Y.; Liu, S.; Wang, F.; Chen, Y. J.; Qian, C.; Feng, J. S. PPDM: Parallel point detection and matching for real-time human–object interaction detection. *arXiv preprint* arXiv:1912.12898, 2020.

[29] He, K. M.; Gkioxari, G.; Dollar, P.; Girshick, R. B. ”Mask R-CNN”. *IEEE Transactions on Pattern Analysis and Machine Intelligence* Vol. 42, No. 2, 386–397, 2020.

[30] Fang, H. S.; Xie, S. Q.; Tai, Y. W.; Lu, C. W. RMPE: Regional multi-person pose estimation. *arXiv preprint* arXiv:1612.00137, 2016.

[31] Fang, H. S.; Cao, J. K.; Tai, Y. W.; Lu, C. W. Pairwise body-part attention for recognizing human–object interactions. In: *Computer Vision—ECCV 2018. Lecture Notes in Computer Science, Vol. 11214.* Ferrari, V.; Hebert, M.; Sminchisescu, C.; Weiss, Y. Eds. Springer Cham, 52–68, 2018.

[32] Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G.; Dean, J. 2013.Distributed representations of words and phrases and their compositionality. In: Proceedings of the 26th International Conference on Neural Information Processing Systems, Vol. 2, 3111–3119, 2013.

[33] Lin, T. Y.; Goyal, P.; Girshick, R.; He, K. M.; Dollár, P. Focal loss for dense object detection. In: Proceedings of the IEEE International Conference on Computer Vision, 2999–3007, 2017.

[34] Lin, T. Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C. L. Microsoft COCO: Common objects in context. In: *Computer Vision—ECCV 2014. Lecture Notes in Computer Science, Vol. 8693.* Fleet, D.; Pajdla, T.; Schiele, B.; Tuytelaars, T. Eds. Springer Cham, 740–755, 2014.

[35] Girshick, R.; Radosavovic, I.; Gkioxari, G.; Dollar, P.; He, K. M. Detectron. 2018. Available at `https://github.com/facebookresearch/detectron`.

[36] Kingma, D. P.; Ba, J. Adam: A method for stochastic optimization. *arXiv preprint* arXiv:1412.6980, 2014.

[37] Zhou, T. F.; Wang, W. G.; Qi, S. Y.; Ling, H. B.; Shen, J. B. Cascaded human–object interaction recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 4262–4271, 2020.

[38] Shen, L.; Yeung, S.; Hoffman, J.; Mori, G.; Fei-Fei, L. Scaling human–object interaction recognition through zero-shot learning. In: Proceedings of the IEEE Winter Conference on Applications of Computer Vision, 1568–1576, 2018.

[39] Li, Y. L.; Liu, X. P.; Lu, H.; Wang, S. Y.; Liu, J. Q.; Li, J. F.; Lu, C. W. Detailed 2D–3D joint representation for human–object interaction. *arXiv preprint* arXiv:2004.08154, 2020.

[40] Li, Y. L.; Xu, L.; Liu, X. P.; Huang, X. J.; Xu, Y.; Wang, S. Y.; Fang, H. S.; Ma, Z.; Chen, M. Y.; Lu, C. W. PaStaNet: Toward human activity knowledge engine. *arXiv preprint* arXiv:2004.00945, 2020.

Hanchao Liu is a master student in the Department of Computer Science and Technology, Tsinghua University. His research interests include image and video processing, and computer vision.

Tai-Jiang Mu is an assistant researcher in the Department of Computer Science and Technology, Tsinghua University, where he received his B.S. and Ph.D. degrees in computer science and technology in 2011 and 2016, respectively. His research interests include visual media learning, SLAM, and human robot interaction.

Xiaolei Huang is an associate professor in the College of Information Sciences and Technology at Pennsylvania State University. Her research interests lie in the intersection of biomedical image analysis, machine learning, and computer vision. She has over 140 publications and holds 7 patents in these areas. She is an associate editor of the *Computer Vision and Image Understanding* journal. She received her bachelor degree in computer science from Tsinghua University, and her master and doctoral degrees in computer science from Rutgers University.