Noname manuscript No.
(will be inserted by the editor)

Polysemy Deciphering Network for Robust Human-Object Interaction Detection

Xubin Zhong[1] · Changxing Ding[1] · Xian Qu[1] · Dacheng Tao[2]

Received: date / Accepted: date

Abstract Human-Object Interaction (HOI) detection is important to human-centric scene understanding tasks. Existing works tend to assume that the same verb has similar visual characteristics in different HOI categories, an approach that ignores the diverse semantic meanings of the verb. To address this issue, in this paper, we propose a novel Polysemy Deciphering Network (PD-Net) that decodes the visual polysemy of verbs for HOI detection in three distinct ways. First, we refine features for HOI detection to be polysemy-aware through the use of two novel modules: namely, Language Prior-guided Channel Attention (LPCA) and Language Prior-based Feature Augmentation (LPFA). LPCA highlights important elements in human and object appearance features for each HOI category to be identified; moreover, LPFA augments human pose and spatial features for HOI detection using language priors, enabling the verb classifiers to receive language hints that reduce intra-class variation for the same verb. Second, we introduce a novel Polysemy-Aware Modal Fusion module (PAMF), which guides PD-Net to make decisions based on feature types deemed more important according to the language priors. Third, we propose to relieve the verb polysemy problem through sharing verb classifiers for semantically similar HOI categories. Furthermore, to expedite research on the verb polysemy problem, we build a new benchmark dataset named HOI-VerbPolysemy (HOI-VP), which includes common verbs (predicates) that have diverse semantic meanings in the real world. Finally, through deciphering the visual polysemy of verbs, our approach is demonstrated to outperform state-of-the-art methods by significant margins on the HICO-DET, V-COCO, and HOI-VP databases. Code and data in this paper will be released at https://github.com/MuchHair/PD-Net.

Keywords Human-object interaction · Verb polysemy · Language priors · Attention model.

✉ Changxing Ding[1]
E-mail: chxding@scut.edu.cn

Xubin Zhong[1]
E-mail: eexubin@mail.scut.edu.cn

Xian Qu[1]
E-mail: eequxian971017@mail.scut.edu.cn

Dacheng Tao[2]
E-mail: dacheng.tao@sydney.edu.au

[1] School of Electronic and Information Engineering, South China University of Technology, Guangzhou 510000, China
[2] UBTECH Sydney Artificial Intelligence Centre and the School of Computer Science, Faculty of Engineering, The University of Sydney, Darlington, NSW 2008, Australia

1 Introduction

In recent years, researchers working in the field of computer vision have begun to pay increasing attention to scene understanding tasks (Choi et al. 2015; Zheng et al. 2015; Lu et al. 2016; Zhang et al. 2017; Zhao et al. 2020; Lin et al. 2020). Since human beings are often central to real-world scenes, Human-Object Interaction (HOI) detection has become a fundamental problem in scene understanding. HOI detection involves not only identifying the classes and locations of objects in the images, but also the interactions (verbs) between each human-object pair. As shown in Fig. 1, an interaction between a human-object pair can be represented by a triplet $<person\ verb\ object>$, herein referred to as one HOI category. One human-object pair may comprise multiple triplets, e.g. $<person\ fly\ airplane>$ and $<person\ ride\ airplane>$.

The HOI detection task is notably challenging (Chao et al. 2018; Gao et al. 2018). One major reason is that verbs can be polysemic. As illustrated in Fig. 1, a verb may convey substantially different semantic meanings and visual characteristics with respect to different objects, as these objects

(a) person play soccer (b) person play frisbee

(c) person hold book (d) person hold elephant

(e) person fly kite (f) person fly airplane

Fig. 1 Examples reflecting the verb polysemy problem in HOI detection. In terms of describing the HOIs, (a) and (b) present HOI examples of "play". "feet" are more important in (a) while "hands" are more important in (b). (c) and (d) illustrate HOI examples of "hold". The human-object pairs in (c) and (d) are characterized by dramatically different human-object spatial features, i.e. the relative location between two bounding boxes. (e) and (f) illustrate HOI examples of "fly". The "person" in (e) exhibits discriminative pose features while the "person" in (f) does not.

may have diverse functions and attributes. One pair of examples can be found in Fig. 1(a) and (b). Here, the "feet" are the more discriminative parts of the human figure for $<person\ play\ soccer>$ while "hands" are more important for describing $<person\ play\ frisbee>$. A second pair of examples is presented in Fig. 1(c) and (d). The human-object pairs in (c) and (d), despite being tagged with the same verb, present dramatically different human-object spatial features (i.e. the relative location between two bounding boxes). Another more serious consideration is that the importance of the same type of visual feature may vary dramatically as the objects of interest change. For example, the human pose plays a vital role in describing $<person\ fly\ kite>$ in Fig. 1(e); by contrast, the human pose is invisible and therefore useless for characterizing $<person\ fly\ airplane>$ in Fig. 1(f). Verb polysemy therefore presents a significant challenge in the HOI detection.

The problem of verb polysemy is relatively underexplored, and sometimes even ignored, in existing works (Xu et al. 2019; Li et al. 2019b; Liao et al. 2020; Wang et al. 2020). Most contemporary approaches tend to assume that the same verb will have similar visual characteristics across different HOI categories, and accordingly opt to design Object-SHared (SH) verb classifiers. When the verb classifier is shared among all objects, each verb obtains more training samples, thereby promoting the robustness of the classification for HOI categories with a small sample size. However, due to the polysemic nature of the verbs, a dramatic semantic gap may exist between instances of the same verb across different HOI categories. Chao *et al.* (2018) constructed Object-SPecific (SP) verb classifiers for each HOI category, which are able to overcome the polysemy problem for HOI categories that have sufficient training samples. However, this approach lacks few- and zero-shot learning abilities for HOI categories where only small amounts of training data are available.

In this paper, we propose a novel Polysemy Deciphering Network (PD-Net) to address the challenging verb polysemy problem. As illustrated in Fig. 2, PD-Net transforms the multi-label verb classifications for each human-object pair into a set of binary classification problems. Here, each binary classifier is used for the verification of one verb category. The classifiers share the majority of their parameters; the main difference lies in the input features. Next, we decode the verb polysemy problem in the following three ways.

First, we enable features sent for each binary classifier to be polysemy-aware using two novel modules, namely Language Prior-guided Channel Attention (LPCA) and Language Prior-based Feature Augmentation (LPFA). The language priors are word embeddings of phrases made up of one verb and one object. The object class is predicted by one object detector; the verb is the one to be determined by one specific binary verb classifier. For its part, LPCA is applied to both the human and object appearance features. The two appearance features are usually redundant, as only part of their information is involved for one specific HOI category (see Fig. 1). Therefore, LPCA is used to highlight important elements in the appearance features for each binary classifier. We further introduce a specific supervision signal for LPCA that enables its parameters to be effectively optimized. Moreover, both human-object spatial and human pose features are often vague and can vary dramatically for the same verb, as shown in Fig. 1(a) and (b), (c) and (d); we therefore propose LPFA, which concatenates the two features with language priors, respectively. In this way, the classifiers can receive hints to reduce the intra-class variation of the same verb for the pose and spatial features.

We further design a novel Polysemy-Aware Modal Fusion module (PAMF), which produces attention scores based on the above language priors in order to dynamically fuse four feature types: human appearance, object appearance, human pose, and spatial features. The language priors pro-

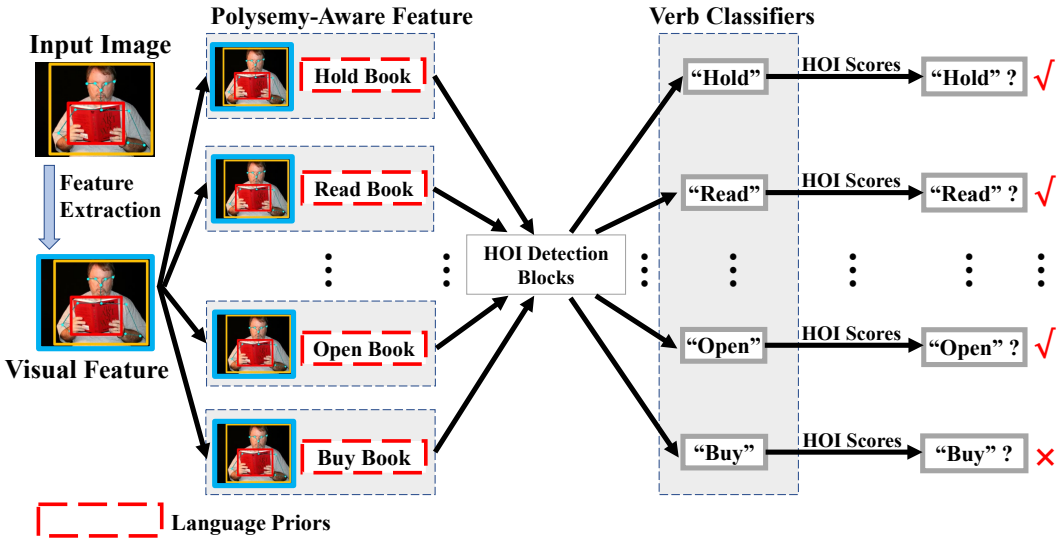

Fig. 2 Visual features of each human-object pair are duplicated multiple times so that polysemy-aware visual features can be obtained under the guidance of language priors. Each polysemy-aware feature is sent to a specific verb classifier. To reduce the number of duplicated human-object pairs, meaningless HOI categories (e.g. $<person\ eat\ book>$ and $<person\ ride\ book>$) are ignored. Meaningful and common HOI categories (e.g. $<person\ hold\ book>$ and $<person\ open\ book>$) are available in each popular HOI detection database.

vide hints regarding the importance of the features for each HOI category. As can be seen in Fig. 1, the human pose feature is discriminative when the language prior is "fly kite" (Fig. 1(e)), but is less useful when the language prior is "fly airplane" (Fig. 1(f)). Therefore, our proposed PAMF deciphers the verb polysemy problem by highlighting the features that are more important for each HOI category.

Moreover, as mentioned above, both SH and SP verb classifiers have limitations. We therefore propose a novel Clustering-based Object-SPecific verb classifier (CSP), which combines the advantages of SH and SP verb classifiers. The main motivation is to ensure that semantically similar HOI categories containing the same verb, e.g. $<person\ hold\ elephant>$, $<person\ hold\ horse>$, and $<person\ hold\ cow>$ can share the same verb classifier. HOI categories that are semantically very different (e.g. $<person\ hold\ book>$ and $<person\ hold\ backpack>$) are identified using another verb classifier. In this way, the verb polysemy problem is mitigated. Meanwhile, CSP has the capacity to handle the few- and zero-shot learning problems that arise in HOI detection, since we merge the training data of semantically similar HOI categories.

To the best of our knowledge, our proposed PD-Net is the first approach to explicitly handle the verb polysemy problem in HOI detection. Moreover, and more impressively, our experimental results on three databases demonstrate that our approach consistently outperforms state-of-the-art methods by considerable margins. A preliminary version of this paper has been published in (Zhong et al. 2020). Compared with the conference version, this version further proposes the novel LPCA module, simplifies the architecture of PD-Net by using CSP classifiers, builds a new database (named HOI-VP) to facilitate the research on the verb polysemy problem, and includes further experimental investigations.

The remainder of this paper is organized as follows. Section 2 briefly reviews related works. The details of the proposed components of PD-Net are described in Section 3. The databases and implementation details are introduced in Section 4, while the experimental results are presented in Section 5. Finally, we conclude the paper in Section 6.

2 Related Works

Human-Object Interaction Detection. HOI detection performs multi-label verb classification for each human-object pair, meaning that the interaction between the same human-object pair may be described using multiple verbs. Depending on the order of verb classification and target object association, existing HOI detection approaches can be divided into two categories. The first category of methods infer the verb actions being performed by one person, then associate each verb with a single object in the image. Multiple target object association approaches have been proposed. For example, Shen *et al.* (2018) proposed an approach based on the value of object detection scores, while Gkioxari *et al.* (2018) fitted a distribution density function of the target object locations based on the human appearance feature. Moreover, Qi *et al.* (2018) adopted a graph parsing network to associate the target objects. Liao *et al.* (2020) and Wang *et al.* (2020) first defined interaction points for HOI detection; next, they locate the interaction points and associate each point with one human-object pair.

The second category of methods first pair each human instance with all object instances as candidate human-object pairs, then recognize the verb for each candidate pair (Gupta et al. 2019). Many types of features have been employed to promote the verb classification performance. For example, Wan *et al.* (2019) employed both human parts and pose-aware features for verb classification, while Xu *et al.* (2020) exploited human gaze and intention to assist HOI detection. Furthermore, Wang *et al.* (2019a) extracted context-aware human and object appearance features to promote HOI detection performance. Li *et al.* (2020a) utilized 3D pose models and 3D object location to assist HOI detection. Moreover, Li *et al.* (2020b) annotated large amounts of part-level human-object interactions and trained a PaStaNet, which is helpful for HOI detection models to make use of fine-grained human part features. A large number of novel model architectures for HOI detection have been also developed. For example, Li *et al.* (2019b) introduced a Transferable Interactiveness Network that suppresses candidate pairs without interactions. Peyre *et al.* (2019) constructed a multi-stream model that projects visual features and word embeddings to a joint space, which is helpful for unseen HOI category detection. Xu *et al.* (2019) constructed a graph neural network to promote the quality of word embeddings by utilizing the correlation between semantically similar verbs, while Zhou *et al.* (2020) proposed a cascade architecture that facilitates coarse-to-fine HOI detection.

The Exploitation of Language Priors. Language priors have also been successfully utilized in many computer vision-related fields, including Scene Graph Generation (Lu et al. 2016; Zhang et al. 2017; Gu et al. 2019; Wang et al. 2019b), Image Captioning (Zhou et al. 2019a; Yao et al. 2019), and Visual Question Answering (Zhou et al. 2019a; Gao et al. 2019; Marino et al. 2019). Moreover, several works (Xu et al. 2019; Peyre et al. 2019) have adopted language priors for HOI detection. All of these approaches project visual features and word embeddings to a joint space, which improves HOI detection by exploiting the semantic relationship between similar verbs or HOI categories (e.g. "drink" and "sip" or "ride horse" and "ride cow"). However, these works do not employ language priors to solve the challenging verb polysemy problem. Compared with the above methods, PD-Net aims to solve the verb polysemy problem by using three novel language prior-based components: namely, LPCA, LPFA, and PAMF.

Attention Models. Attention mechanisms are becoming a popular component in computer vision tasks, including Image Captioning (Chen et al. 2017; Xu et al. 2015; You et al. 2016), Action Recognition (Girdhar and Ramanan 2017; Meng et al. 2019), and Pose Estimation (Li et al. 2019a; Ye et al. 2016). Existing studies on attention mechanisms can be roughly divided into three categories: namely, hard regional attention (Jaderberg et al. 2015; Li et al. 2018), soft spatial attention (Wang et al. 2017; Pereira et al. 2019; Zhu et al. 2018), and channel attention (Pereira et al. 2019; Hu et al. 2018). Hard regional attention methods typically predict regions of interest (ROIs) first, and then only utilize features in ROIs for subsequent tasks. In comparison, soft spatial attention and channel attention (CA) models use soft weights to highlight important features in the spatial and channel dimensions, respectively. There have also been existing works that adopted attention models to assist with HOI detection tasks. For example, Gao *et al.* (2018) and Wang *et al.* (2019a) employed an attention mechanism to enhance the human and object features by aggregating contextual information. Wan *et al.* (2019) proposed the PMFNet model, which adopts human pose and spatial features as cues to infer the importance of each human part. Ulutan *et al.* (2020) used cues derived from a human-object spatial configuration to highlight the important elements in appearance features.

To the best of our knowledge, few works thus far have made use of the attention mechanism to solve the verb polysemy problem in HOI detection. Moreover, existing attention models for HOI detection usually employ visual features (e.g. the appearance and pose features) as cues. By contrast, our proposed LPCA and PAMF adopt language priors as cues; these priors have clear semantic meanings, and are therefore well-suited to resolving the verb polysemy problem.

3 Method

3.1 Overview

The framework of PD-Net is illustrated in Fig. 3. Once given an image, human and object proposals are generated using Faster R-CNN (Ren et al. 2015). Each human proposal h and each object proposal o are paired as a candidate for verb classification. PD-Net produces a set of verb classification scores for each candidate pair. Similar to existing works (Li et al. 2019b; Gupta et al. 2019; Wan et al. 2019; Zhou and Chi 2019; Li et al. 2020a), four types of visual features are adopted, i.e., human appearance, object appearance, human-object spatial, and human pose features. The human and object appearance features are extracted from Faster R-CNN model using human and object bounding boxes, respectively. The human-object spatial feature is a 42-dimensional vector encoded using the bounding box coordinates of one human-object pair following (Gupta et al. 2019). Moreover, we use a pose estimation model (Fang et al. 2017) to obtain the coordinates of 17 keypoints for each human instance. Subsequently, following (Gupta et al. 2019), the human keypoints and the bounding box coordinates of object proposal are then encoded into a 272-dimensional pose feature vector.

As outlined in Fig. 2, we transform the multi-label verb classification into a set of binary classification problems.

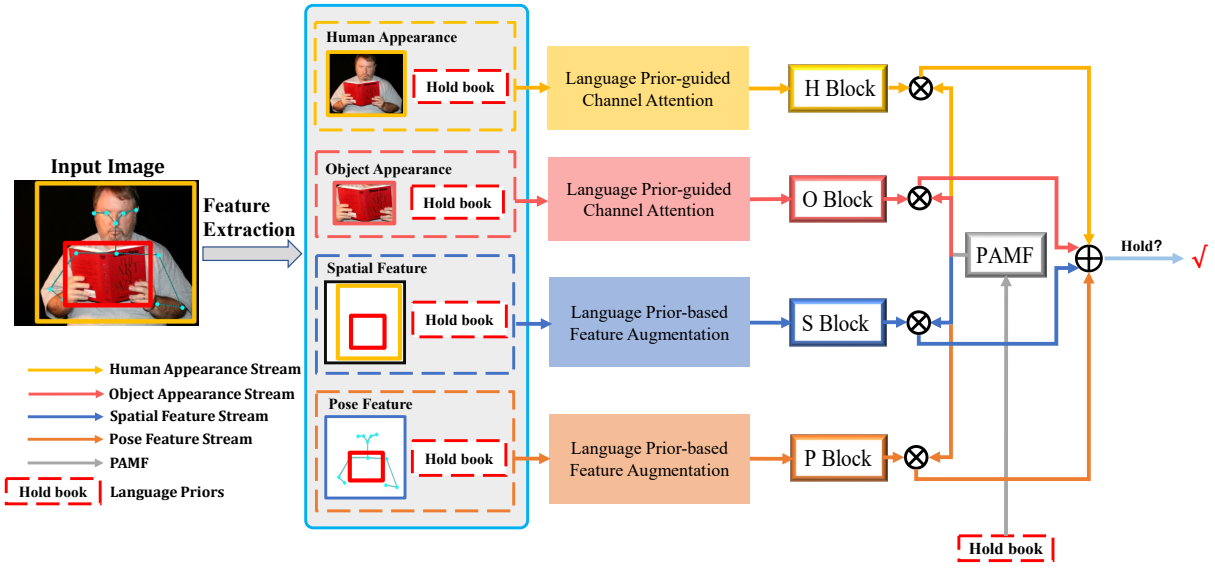

Fig. 3 Overview of the Polysemy Deciphering Network. For the sake of simplicity, only one binary CSP classifier (for "hold") is illustrated here. PD-Net takes four feature streams as input: the human appearance stream (**H** stream), the object appearance stream (**O** stream), the human-object spatial stream (**S** stream), and the human pose stream (**P** stream). These four feature streams are first processed by either LPCA or LPFA to be polysemy-aware. They are then sent to the **H**, **O**, **S**, and **P** blocks for binary classification, respectively. Subsequently, the classification scores from the four feature streams are fused using the attention scores produced by **PAMF**. Here, $\otimes$ and $\oplus$ denote the element-wise multiplication and addition operations, respectively.

Each of the binary classifiers is used for one verb category verification and includes a set of **H**, **O**, **S**, and **P** blocks. Apart from the final layer which is used for verb prediction, the parameters of the other layers in each respective block are shared across different binary classifiers. Therefore, their overall model size is comparable to an ordinary multi-label classifier. The binary classifiers mainly differ in terms of their input features and the way in which they combine predictions from the four feature streams. In the following, we propose four novel components to handle the verb polysemy problem in HOI detection. First, we introduce the LPCA and LPFA modules, which facilitate the four types of features being polysemy-aware. Second, we design the PAMF module that adaptively fuses the prediction scores produced by the four feature streams and obtains the final prediction score for each binary classifier. Finally, we propose a CSP classification scheme that strikes a balance between resolving the verb polysemy problem and reducing the number of binary classifiers in PD-Net.

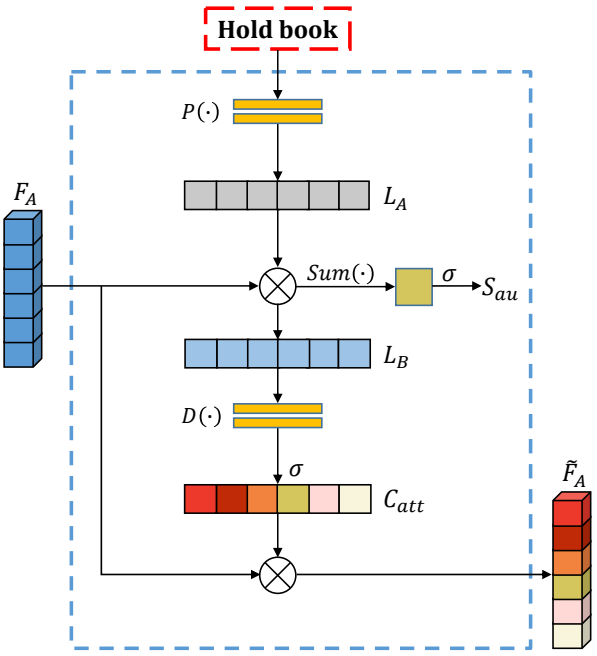

Fig. 4 Model structure of LPCA. F_A denotes the human or object appearance feature. $P(\cdot)$ and $D(\cdot)$ are realized by means of multi-layer perceptrons (MLPs). $\otimes$ denotes the element-wise multiplication operation.

3.2 Polysemy-Aware Feature Generation

We here introduce two novel components, i.e. LPCA and LPFA, that generate polysemy-aware features with the help of language priors. The language prior we used in this paper is the concatenated word embedding of two words: one word denotes the verb to be identified, while the other one is the detected object in the human-object pair. The word embeddings are generated using the word2vec tool (Mikolov et al. 2013). The dimension of the language prior is 600.

Language Prior-guided Channel Attention. Both the human and object appearance features are usually redundant as

only part of their information is involved for a specific HOI category. One example can be found in Fig. 1(a) and (b), where the human body parts most relevant to "play soccer" and "play frisbee" are the "feet" and "hands", respectively. We therefore propose LPCA to highlight the important elements in human and object appearance features, based on the channel attention scheme and the guidance of language priors. LPCA is realized through three steps, which are outlined below.

First, we infer the important elements in the human or object appearance feature F_A using language priors. As can be seen from Fig. 4, the language prior is projected to a K_A-dimensional vector (denoted as L_A) via an MLP. K_A is equal to the dimension of F_A. The MLP is realized by two successive FC layers and the dimension of the first one is set to $\frac{K_A}{2}$. Similar to (Peyre et al. 2019), L_A is normalized via its L2 norm. To drive L_A to pay attention to important elements in F_A regarding to the verb-object pair, we perform element-wise multiplication between L_A and F_A, as follows:

$$L_B = F_A \otimes L_A, \tag{1}$$

and further compute the summation of elements in L_B:

$$\mathcal{S}_{au} = \sigma(Sum(L_B)), \tag{2}$$

where σ denotes the sigmoid activation function and $Sum(\cdot)$ denotes the summation of all elements in one vector. During the training stage, we minimize the binary cross-entropy loss between $\mathcal{S}_{au}$ and the binary label for the verb to verify. During inference, the operation in Eq. (2) can be ignored. By optimizing the FC layers via the verb verification goal, the value of elements in L_A can reflect the importance of the corresponding elements in F_A.

It is worth noting that the quality of both L_A and L_B can be affected by the discrepancy between visual features and word embeddings, since the word embeddings are not specifically designed for computer vision tasks (Xu et al. 2019). Therefore, directly using L_B as representation for verb verification may be suboptimal. Consequently, to handle this problem, we propose the following strategy to further enhance the quality of the representations.

Second, we obtain attention scores based on L_B via a plain CA module:

$$C_{att} = \sigma(D(L_B)), \tag{3}$$

where C_{att} stands for the final channel attention scores for F_A. $D(\cdot)$ is another MLP realized by two successive FC layers. For its part, the plain CA module makes use of the correlation between elements in L_B to promote the quality of attention scores.

Finally, the polysemy-aware human or object appearance features can be obtained via the following equation:

$$\tilde{F}_A = F_A \otimes C_{att}. \tag{4}$$

Unlike existing CA models, e.g. the Squeeze and Excitation (SE) network (Hu et al. 2018) and VSGNet (Ulutan et al. 2020), LPCA includes an explicit supervision signal that guides PD-Net to generate polysemy-aware features. Moreover, VSGNet adopts human-object spatial configuration as cues to infer channel attention. However, the spatial configuration may be vague to precisely infer important elements in F_A. In comparison, the language priors adopted in LPCA contain clearer semantic information for each HOI. Moreover, language priors are adopted in (Peyre et al. 2019) to construct verb classifiers. In their method, human and object features are fixed. In comparison, we utilize language priors to generate polysemy-aware features, which are adjustable for each verb-object pair to identify.

Language Prior-based Feature Augmentation. LPFA is applied to human-object spatial and human pose features. As illustrated in Fig. 1(c) and (d), (e) and (f), the spatial and pose features are often vague and vary dramatically for the same verb, meaning that they contain insufficient information. Therefore, we propose LPFA to augment the pose and spatial features. More specifically, we concatenate each of the two features with the 600-dimensional language prior. As a result of this concatenation, the classifiers receive hints that can aid in reducing the intra-class variation of the same verb for the pose and spatial features.

3.3 Polysemy-Aware Modal Fusion

As illustrated in Fig. 3, the four feature streams are sent to the **H**, **O**, **S** and **P** blocks, respectively. The **H** and **O** blocks are constructed using two successive FC layers, while the **S** and **P** blocks are constructed using three successive FC layers. In the interests of simplicity, the dimension of each hidden FC layer is set as the dimension of its input feature vector. The output dimension of these four blocks is set to K_C, which is the number of binary classifiers in PD-Net.

As discussed in Section 1, one major challenge posed by the verb polysemy problem is that the relative importance of each of the four feature streams to the identification of the same verb may vary dramatically as the objects change. As shown in Fig. 1(e) and (f), the human appearance and pose features are most important for detecting $<person\ fly\ kite>$; by contrast, these features are almost invisible and therefore less useful for detecting $<person\ fly\ airplane>$. Therefore, we propose PAMF to generate attention scores that dynamically fuse the predictions of the four feature streams. In more detail, we use the same 600-dimensional word embedding (e.g. "hold book") that is implemented in LPCA and LPFA. The language prior is fed into two successive FC layers, the dimensions of which are 48 and 4, respectively. The first FC layer is followed by a ReLU layer, while the second one is followed by a sigmoid activation function. The output of PAMF is used as attention scores for the four feature

streams. In this way, the important feature streams with respect to each HOI category is highlighted, while those that are less important are suppressed.

Given one human-object pair (h, o), the identification score for one specific verb v obtained by the corresponding binary classifier can be denoted as $\mathcal{S}^{\mathbf{PD}}_{(h,o,v)}$:

$$\mathcal{S}^{\mathbf{PD}}_{(h,o,v)} = \sigma(\sum_{i\in\{\mathbf{H},\mathbf{O},\mathbf{S},\mathbf{P}\}} a_{(i,o,v)} s_{(i,o,v)}), \tag{5}$$

where i denotes one feature stream, while $a_{(i,o,v)}$ is the attention score generated by PAMF for the i-th feature stream. $s_{(i,o,v)}$ is the verb prediction score generated by the i-th feature stream.

3.4 Clustering-based Object Specific Verb Classifiers

Although LPCA, LPFA, and PAMF can help SH verb classifiers to relieve the verb polysemy problem, the defects of SH verb classifiers still remains. The essential reason is that HOIs with different objects share the same verb classifier. Under ideal circumstances, SP verb classifiers can overcome this problem if sufficient training data exists for each HOI category. However, if we assume that the number of object categories is $|O|$ and the number of verb categories is $|V|$, the total number of their combinations will therefore be $|O| \times |V|$, which is usually very large even if meaningless HOI categories are removed. Therefore, it is too difficult to obtain sufficient train samples for each HOI category. Moreover, due to the class imbalance problem for HOI categories, the SP classifiers lack few- and zero-shot learning abilities for HOI categories which have small amount of training data. Therefore, both types of verb classifiers have limitations.

In this subsection, we introduce a novel verb classifier, named Clustering-based object-SPecific (CSP) verb classifiers. CSP classifiers can strike a balance between overcoming the verb polysemy problem and handling the zero- or few-shot learning problems. The main motivation behind CSP classifiers is that some HOIs tagged with the same verb are both semantically and visually similar, e.g. $<person\ hold\ sheep>$, $<person\ hold\ horse>$, and $<person\ hold\ cow>$; therefore, they can share the same verb classifier, meaning that the number of SP classifiers is consequently reduced. In more detail, we first obtain all meaningful and common HOI categories for each verb, which are available in popular databases such as HICO-DET (Chao et al. 2018) and V-COCO (Gupta and Malik 2015). The number of meaningful HOI categories including the verb v is indicated by O_v. We then use the K-means method (MacQueen et al. 1967) to cluster the HOI categories with the same verb v into C_v clusters according to the cosine distance between the word embeddings of the objects. We empirically set the C_v for each verb as a rounded number of the square root of O_v. This clustering strategy is also capable of handling the few- and zero-shot learning problems of SP verb classifiers. For example, during testing, a new HOI category $<person\ hold\ elephant>$ can share the same classifier with other HOI categories that have similar semantic meanings (e.g. $<person\ hold\ horse>$).

3.5 Training and Testing

Training. PD-Net can be conceptualized as a multi-task network. Its loss for the verification of the verb v in one HOI category (h, v, o) can be represented as follows:

$$\mathcal{L}_{(h,o,v)} = \mathcal{L}_{BCE}(\mathcal{S}^{\mathbf{PD}}_{(h,o,v)}, l_v) + \sum_{i\in\{\mathbf{H},\mathbf{O}\}} \mathcal{L}_{BCE}(\mathcal{S}^{i}_{au}, l_v), \tag{6}$$

where $\mathcal{L}_{BCE}$ represents binary cross-entropy loss, while l_v denotes a binary label ($l_v \in \{0, 1\}$) for one verb to verify. Moreover, $\mathcal{S}^{\mathbf{H}}_{au}$ and $\mathcal{S}^{\mathbf{O}}_{au}$ denote the output of Eq. (2) for the human and object appearance features, respectively.

Testing. During testing, we use the same method as that utilized in the training stage to obtain the language priors. Here, the object category in the prior is predicted using Faster R-CNN (rather than the ground-truth); the verb category in the prior varies for each binary classifier of the verb. Following (Li et al. 2019b, 2020a; Ulutan et al. 2020; Wan et al. 2019), we also construct an Interactiveness Network (INet) capable of suppressing pairs without interaction. Finally, the prediction score for one HOI category (h, v, o) is represented as follows:

$$\mathcal{S}^{\mathbf{HOI}}_{(h,o,v)} = \mathcal{S}_h \times \mathcal{S}_o \times \mathcal{S}^{\mathbf{PD}}_{(h,o,v)} \times \mathcal{S}^{\mathbf{I}}_{(h,o)}, \tag{7}$$

where $\mathcal{S}_h$ and $\mathcal{S}_o$ are the detection scores of human and object proposals, respectively, while $\mathcal{S}^{\mathbf{I}}_{(h,o)}$ denotes the prediction score generated by the pre-trained INet. In the experimental section below, we demonstrate that INet slightly promotes the performance of PD-Net.

4 Experimental Setup

4.1 Datasets

HICO-DET (Chao et al. 2018) is a large-scale dataset for HOI detection, containing a total of 47,776 images; of these, 38,118 images are assigned to the training set, while the remaining 9,568 images are used as the testing set. There are 117 verb categories, 80 object categories, and 600 common HOI categories overall; moreover, these 600 HOI categories are divided into 138 rare and 462 non-rare categories. Each rare HOI category contains less than 10 training samples. Each verb is included in an average of five HOI categories.

Table 1 The number of associated objects and instances for each verb in the HOI-VP dataset.

Verbs	# Objects	# Instances
carry	49	1585
cross	8	611
fix	4	2063
hold	229	21502
in	218	6123
on	196	26427
make	8	147
open	7	181
operate	4	46
play	22	1119
push	10	202
ride	29	1734
swing	5	1116
touch	18	154
use	29	3333

V-COCO (Gupta and Malik 2015) is a subset of MS-COCO (Lin et al. 2014) and contains 2,533, 2,867, and 4,946 images used for training, validation and testing, respectively. There are 24 verb categories and 259 HOI categories in total. Each verb is included in 10 HOI categories on average.

HOI-VerbPolysemy (HOI-VP) is a new database constructed in this paper. To the best of our knowledge, this is the first database to be designed explicitly for the verb polysemy problem in HOI detection. In more detail, it consists of 15 common verbs (predicates) that have rich and diverse semantic meanings. It also contains 517 common objects in real-world scenarios. Each verb is included in an average of 55 HOI categories, as detailed in Table 1. In particular, "in" and "on" are two highly common predicates that are also polysemic in visual relationship detection tasks (Lu et al. 2016; Krishna et al. 2017; Kuznetsova et al. 2020; Ji et al. 2020) and are thus both included in the HOI-VP database. There are 21,928 and 7,262 images used for training and testing, respectively. All images are collected from the VG database (Krishna et al. 2017), while the corresponding annotations are provided by the HCVRD database (Zhuang et al. 2017). It is worth noting here that the annotations in HCVRD contain noise. For example, the same verb may be annotated with different words, e.g., "hold", "holds", and "holding", while a similar problem exists for the objects, e.g., "camera", "digital camera", and "video camera". We therefore merge different annotations for the same verb or object categories, respectively. Some sample images from HOI-VP are illustrated in Fig. 9. This database will be made publicly available to expedite research into the verb polysemy problem.

4.2 Evaluation Metrics

According to the official protocols (Chao et al. 2018; Gupta and Malik 2015), mean average precision (mAP) is used as the evaluation metric for HOI detection on both HICO-DET and V-COCO datasets. A positive human-object pair must meet the following requirements: first, the predicted HOI category must be the same type as the ground truth; second, both the human and object proposals must have an Intersection over Union (IoU) with the ground truth proposals of more than 0.5. Moreover, there are two mAP modes in HICO-DET, namely the **Default** (DT) mode and the **Known-Object** (KO) mode. In the DT mode, we calculate the average precision (AP) for each HOI category in all testing images. In the KO mode, the object categories in all images are known; therefore, we need only to compute the AP for each HOI category from images containing the interested object. For example, we evaluate the AP of $<person\ ride\ horse>$ using only those testing images that contain a "horse". Since the images that contain the object category of interest are known, the KO mode is better able to reflect the verb classification ability. For V-COCO, the role mAP (Gupta and Malik 2015) (AP_{role}) is used for evaluation.

Table 2 Performance comparisons between common object detectors on COCO (Lin et al. 2014).

Method	Backbone	AP
Faster R-CNN (Ren et al. 2015)	ResNet-152	36.7 (Chen and Gupta 2017)
Faster R-CNN	ResNet-50-FPN	36.8 (Massa and Girshick 2018)
Mask R-CNN (He et al. 2017)	ResNet-50	36.9 (Girshick et al. 2018)
CenterNet (Zhou et al. 2019b)	Hourglass-104 (Newell et al. 2016)	40.3 (Zhou et al. 2019b)
Faster R-CNN	NASNet (Zoph et al. 2018)	43.0 (Huang et al. 2017)

For the HOI-VP database, we use an evaluation protocol similar to that of HICO-DET. As there are as many as 517 object categories in HOI-VP, object detection becomes a challenging task. Accordingly, to reduce the impact of object detection errors, the ground-truth bounding boxes and categories for both human and object instances are provided. This strategy is similar to the Predicate Classification (PREDCLS) protocol, which has been widely adopted in scene graph generation tasks (Zellers et al. 2018; Lin et al. 2020). It facilitates a clean comparison of verb classification ability between different HOI detection models.

4.3 Implementation Details

To facilitate fair comparison with existing works, we consider two popular object detection models for PD-Net. The first of these, Faster R-CNN (Ren et al. 2015) with ResNet-50-FPN (Lin et al. 2017) backbone, attaches a Feature Pyramid Network (FPN) to ResNet-50 (He et al. 2016) and generates object proposals from the FPN. Based on these proposals, instance appearance features are extracted from the ResNet-50 model. The second model is Faster R-CNN with ResNet-152 backbone (He et al. 2016). Here, both instance proposals and appearance features are obtained from the ResNet-152 model. The above two object detectors are trained on the COCO database (Lin et al. 2014). As shown in Table 2,

Table 3 Ablation studies on each component of PD-Net. Full refers to evaluation on all 600 HOI categories in HICO-DET.

Methods	Components						Full	
	SH	PAMF	LPFA	LPCA	CSP	INet	DT	KO
Our Baseline	✓	-	-	-	-	-	17.57	23.07
Incremental	✓	✓	-	-	-	-	18.86	24.43
	✓	✓	✓	-	-	-	19.38	24.64
	✓	✓	✓	✓	-	-	20.71	24.85
	-	✓	✓	✓	✓	-	21.77	**26.98**
Drop-one-out	-	-	✓	✓	✓	-	20.14	25.65
	-	✓	-	✓	✓	-	21.37	26.15
	-	✓	✓	-	✓	-	19.32	24.30
	✓	✓	✓	✓	-	-	20.71	24.85
PD-Net	-	✓	✓	✓	✓	✓	**22.37**	26.86

the two object detectors achieve comparable detection performance. Moreover, to facilitate fair comparison with the majority of existing works (Ulutan et al. 2020; Gupta et al. 2019; Li et al. 2019b; Peyre et al. 2019; Qi et al. 2018), we fix the parameters of both object detectors. The dimension of appearance features for both object detectors, i.e. K_A, is 2,048.

Utilizing the same approach as existing works (Gao et al. 2018; Xu et al. 2019; Li et al. 2019b; Gupta et al. 2019; Peyre et al. 2019; Qi et al. 2018; Liao et al. 2020; Wang et al. 2020; Ulutan et al. 2020), the HOI categories that appear in the training set are set as the meaningful and common HOI categories in each HOI database. The dimension of output layers of the **H**, **O**, **S**, and **P** blocks, i.e. K_C, is set to 187, 45, and 83 on the HICO-DET, V-COCO, and HOI-VP databases, respectively; these figures are equal to the number of CSP classifiers on each respective database. We train PD-Net for 6 (10) epochs using Adam optimizer (Kingma and Ba 2014) with a learning rate of 1e-3 (1e-4) on HICO-DET (V-COCO), while on HOI-VP, we train PD-Net for 12 epochs using a learning rate of 1e-3. During testing, we rank the HOI candidate pairs according to their detection scores (obtained via Eq. (7)) and calculate mAP for evaluation purposes.

5 Experimental Results and Discussion

5.1 Ablation Studies

To demonstrate the effectiveness of each proposed component in PD-Net, we perform ablation studies on the HICO-DET database. In Table 3, the baseline is constructed by removing LPCA, LPFA, and PAMF from PD-Net; we also replace CSP classifiers with SH classifiers. The other settings for the baseline remain the same as in PD-Net. For both models, the Faster R-CNN with ResNet-152 backbone is used for object detection. Experimental results are summarized in Table 3. From these results, we can make the following observations.

Effectiveness of PAMF. PAMF is designed to decipher the verb polysemy by assigning larger weights to more important feature types for each HOI category. As shown in Table 3, PAMF promotes the performance of the baseline by 1.29% and 1.36% mAP in DT and KO modes, respectively.

Effectiveness of LPFA. LPFA is used to provide hints for the classifier in order to reduce the intra-class variation of the pose and spatial features by augmenting them with language priors. When LPFA is incorporated, HOI detection performance is promoted by 0.52% and 0.21% mAP in DT and KO modes, respectively.

Effectiveness of LPCA. The appearance features are redundant for HOI detection. LPCA is proposed to generate polysemy-aware appearance features. As can be seen from Table 3, LPCA promotes the HOI detection performance by a clear margin of 1.33% and 0.21% in DT and KO modes, respectively.

Effectiveness of CSP Classifiers. CSP classifiers can relieve the verb polysemy problem by assigning the same verb classifier to semantically similar HOI categories. As shown in Table 3, CSP classifiers improve the HOI detection performance by 1.06% and 2.13% mAP in DT and KO modes, respectively.

Drop-one-out Study. We further perform a drop-one-out study in which each proposed component is removed individually. These experimental results further demonstrate that each component is indeed helpful to promote HOI detection performance.

Finally, when INet is integrated, the mAP of PD-Net in the DT mode is further promoted by 0.60%. However, the mAP in the KO mode does not improve. This is because INet can assist PD-Net by suppressing candidate pairs without interactions, which are usually caused by incorrect or redundant object proposals in the DT mode. However, the KO mode is comparatively less affected by object detection errors; therefore, PD-Net can achieve high performance without the assistance of INet in this mode. This experiment demonstrates that the strong performance of PD-Net is primarily a result of its excellent verb classification ability.

Table 4 Comparisons with one variant of the language prior.

Language Prior	DT (Full)	KO (Full)
Verb Only	19.98	24.99
Verb + Object	**22.37**	**26.86**

5.2 Comparisons with Variants of PD-Net

5.2.1 Comparisons with Variants of the Language Prior

In this experiment, we remove the word embedding of the object category from the language prior so that only the word embedding of the verb category to identify is used as input for PAMF, LPFA, and LPCA. As shown in Table 4, without the word embedding of the object category, the

Table 5 Performance comparisons with variants for LPCA.

Methods	DT (Full)	KO (Full)
w/o LPCA	19.82	23.94
Plain CA	19.97	24.11
w/o $\mathcal{S}_{au}$	19.41	23.32
w/o C_{att}	21.99	26.34
LPCA	**22.37**	**26.86**

performance of PD-Net drops by a large margin of 2.39% (1.87%) mAP in DT (KO) mode. These experimental results indicate that the word embedding of the object category in the language prior is an important hint to decipher the verb polysemy problem.

5.2.2 Comparisons with Variants of LPCA

In this experiment, we compare the performance of LPCA with three possible variants: namely, Plain CA, 'w/o $\mathcal{S}_{au}$', and 'w/o $\mathcal{C}_{att}$'. The other implementation details of PD-Net are kept the same for different variants. Plain CA means that we feed the appearance feature F_A directly into a plain CA module, i.e. $D(\cdot)$ in Fig. 4, and obtain $\tilde{F}_A$. 'w/o S_{au}' involves removing the extra supervision signal S_{au} from LPCA, while 'w/o C_{att}' means that we directly use L_B in Fig. 4 as the input of the **H** and **O** blocks in Fig. 3, without the further processing by the plain CA module. Experimental results are tabulated in Table 5. In this table, 'w/o LPCA' is a baseline that removes the entire LPCA module from PD-Net. From these results, we can make the following observations.

First, the plain CA module alone slightly promotes the performance of PD-Net. One main reason for this is that the plain CA module has very little ability to identify important elements in the appearance features for each HOI category.

Second, without supervision from S_{au}, the performance of LPCA degrades dramatically. Compared to the plain CA module, this setting adopts language priors to provide cues regarding the channel-wise importance of F_A for each HOI category. However, it receives only implicit supervision from the binary score $\mathcal{S}^{\mathbf{PD}}_{(h,o,v)}$ in Eq. (7), which is too weak to optimize LPCA's parameters. We therefore observe degraded performance after the extra supervision $\mathcal{S}_{au}$ is removed.

Third, 'w/o C_{att}' obtains better performance than both the 'Plain CA' and 'w/o S_{au}' settings. However, its performance is still lower than that of our proposed LPCA by 0.38% and 0.52% mAP in DT and KO modes, respectively. This may be because L_A is obtained via projection from the language prior. As word embeddings are not specifically designed for computer vision tasks, L_A may not always be reliable and the quality of L_B is affected (Xu et al. 2019). Therefore, further processing L_B using the plain CA module is helpful.

Table 6 Performance comparisons between SH, SP, and CSP verb classifiers.

	DT Mode			KO Mode		
Methods	Full	Rare	Non-Rare	Full	Rare	Non-Rare
SH	21.06	16.45	22.43	24.83	19.89	26.30
SP	20.91	15.03	22.66	24.67	17.83	26.71
CSP	**22.37**	**17.61**	**23.79**	**26.86**	**21.70**	**28.44**

In comparison, our proposed LPCA achieves the best performance for the following three reasons. First, it adopts language priors to provide hints regarding the channel-wise importance of F_A for each HOI category. Second, it imposes direct supervision to the attention module, which helps to more effectively optimize the model parameters. Third, it refines the attention vector obtained from the language priors using a plain CA module, which enhances the quality of the channel attention vectors. The above experimental results and analysis demonstrate the effectiveness of LPCA.

5.2.3 Comparisons with Variants of Verb Classifiers

To further demonstrate the advantages of CSP classifiers, we compare their performance with that of SH and SP verb classifiers. To facilitate fair comparison, the other settings of PD-Net remain unchanged. Experimental results are tabulated in Table 6.

It is shown that SH classifiers outperform SP classifiers by 1.42% (2.06%) mAP in DT (KO) mode for rare HOI categories. This is because SH classifiers enable these rare HOI categories to share verb classifiers with other HOI categories that have sufficient training data. By comparison, SP classifiers are better able to relieve the verb polysemy problem for the HOI categories that have sufficient training data. Therefore, the SP classifiers outperform SH classifiers by 0.23% (0.41%) mAP in DT (KO) mode for non-rare HOI categories.

In comparison, CSP classifiers achieve superior performance on both rare and non-rare HOI categories. This is due to the same verb classifiers being assigned to semantically similar HOI categories, enabling HOI categories with few training samples to share verb classifiers with those HOI categories that have sufficient training data. Moreover, different verb classifiers are adopted for semantically different HOI categories, which is helpful to overcome the verb polysemy problem. Overall, CSP classifiers outperform SH and SP classifiers by 1.31% (2.03%) and 1.46% (2.19%) mAP in DT (KO) mode for the full HOI categories, respectively.

5.3 Comparisons with State-of-the-Art Methods

We compare the performance of PD-Net with state-of-the-art methods on three databases, namely HICO-DET, V-COCO,

Table 7 Performance Comparisons on HICO-DET.

Methods	Object Detector Backbone	DT Mode			KO Mode		
		Full	Rare	Non-Rare	Full	Rare	Non-Rare
Shen *et al.* (Shen et al. 2018)	VGG-19	6.46	4.24	7.12	-	-	-
InteractNet (Gkioxari et al. 2018)	ResNet-50-FPN	9.94	7.16	10.77	-	-	-
GPNN (Xu et al. 2019; Qi et al. 2018)	ResNet-152	13.11	9.34	14.23	-	-	-
iHOI (Xu et al. 2020)	ResNet-50-FPN	13.39	9.51	14.55	-	-	-
Xu *et al.* (Xu et al. 2019)	ResNet-50-FPN	14.70	13.26	15.13	-	-	-
iCAN (Gao et al. 2018)	ResNet-50-FPN	14.84	10.45	16.15	16.26	11.33	17.73
Wang *et al.* (Wang et al. 2019a)	ResNet-50-FPN	16.24	11.16	17.75	17.73	12.78	19.21
No-Frills (Gupta et al. 2019)	ResNet-152	17.18	12.17	18.68	-	-	-
TIN (Li et al. 2019b)	ResNet-50-FPN	17.22	13.51	18.32	19.38	15.38	20.57
RPNN (Zhou and Chi 2019)	ResNet-50 (Mask R-CNN)	17.35	12.78	18.71	-	-	-
PMFNet (Wan et al. 2019)	ResNet-50-FPN	17.46	15.65	18.00	20.34	17.47	21.20
Peyre *et al.* (Peyre et al. 2019)	ResNet-50-FPN	19.40	14.60	20.90	-	-	-
IP-Net (Wang et al. 2020)	ResNet-50-FPN	19.56	12.79	21.58	22.05	15.77	23.92
VSGNet (Ulutan et al. 2020)	NASNet (Zoph et al. 2018)	19.80	16.05	20.91	-	-	-
2D-RN ($\mathcal{S}^{2D}$) (Li et al. 2020a)	ResNet-50-FPN	19.98	16.97	20.88	22.56	19.48	23.48
PPDM (Liao et al. 2020)	Hourglass-104	21.73	13.78	24.10	24.58	16.65	26.84
Our baseline	ResNet-50-FPN	17.27	12.27	18.77	23.07	18.29	24.50
PD-Net	ResNet-50-FPN	**20.76**	15.68	**22.28**	**25.59**	**19.93**	**27.28**
Our baseline	ResNet-152	17.57	12.67	19.04	23.07	17.45	24.75
PD-Net	ResNet-152	**22.37**	**17.61**	23.79	**26.86**	**21.70**	**28.44**

and HOI-VP. Experimental results are summarized in Table 7, Table 8, and Table 10, respectively.

5.3.1 Performance comparisons on HICO-DET

As shown in Table 7, PD-Net outperforms state-of-the-art methods by significant margins using both object detector backbones. It is worth noting that one most recent method, i.e. PPDM, adopts CenterNet with Hourglass-104 backbone (Zhou et al. 2019b) as the object detector. As shown in Table 2, this object detector significantly outperforms the two Faster R-CNN object detectors utilized in our model. To facilitate fair comparison, we mainly compare PPDM with PD-Net in the KO mode, as this mode is less affected by object detection results. As shown in Table 7, PD-Net outperforms PPDM in KO mode by significant margins of 2.28%, 5.05% and 1.60% mAP on the full, rare and non-rare HOI categories, respectively. Moreover, PD-Net also outperforms PPDM by 0.64% in the DT mode on the full HOI categories.

Moreover, as shown in Table 7 and Table 2, the object detector adopted by another recent work (Ulutan et al. 2020) is also much stronger than ours. But PD-Net still outperforms this model by large margins of 2.57% (22.37%-19.80%), 1.56% (17.61%-16.05%), and 2.88% (23.79%-20.91%) mAP in the DT mode on the full, rare, and non-rare HOI categories, respectively.

Finally, with a similar multi-stream representation network and object detector backbone (ResNet-50-FPN), PD-Net outperforms one very recent model 2D-RN (Li et al. 2020a) by 3.03% (25.59%-22.56%) and 0.78% (20.76%-19.98%) in mAP on the full HOI categories in the KO and DT modes, respectively. Another advantage of PD-Net compared with 2D-RN is that PD-Net requires no extra human annotation. Besides, 3D human pose and 3D object locations are also utilized to improve 2D-RN during inference in (Li et al. 2020a). To facilitate fair comparisons, we only compare the performance of PD-Net with methods that utilize 2D human pose and 2D object locations during inference.

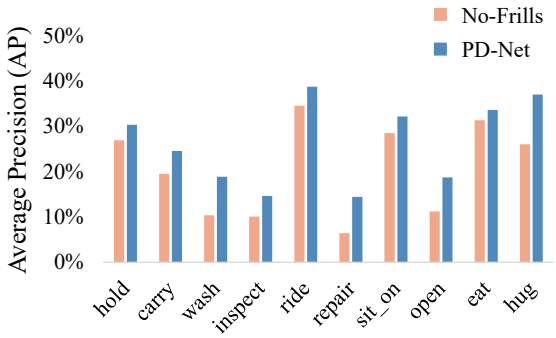

Fig. 5 The top 10 verbs that are most likely to suffer from the polysemy problem on HICO-DET.

To further illustrate the advantage of PD-Net in deciphering the verb polysemy problem, we present the top 10 verbs (from the total 117 verbs in HICO-DET) ranked by the number of HOI categories in which each verb is included in Fig. 5. The largest number of HOI categories associated with the same verb ("hold") is 61. As these verbs are more likely to be affected by the visual polysemy problem, we therefore compare the performance of PD-Net with one state-of-the-art method (Gupta et al. 2019) on these verbs. This method

Table 8 Performance comparisons on V-COCO (Gupta and Malik 2015). ° denotes methods that we reproduce. † denotes methods that adopt human part features as input.

Methods	Object Detector Backbone	AP_{role}
Gupta *et al.* (Gupta and Malik 2015; Gkioxari et al. 2018)	ResNet-50-FPN	31.8
InteractNet (Gkioxari et al. 2018)	ResNet-50-FPN	40.0
GPNN (Xu et al. 2019; Qi et al. 2018)	ResNet-152	44.0
iCAN (Gao et al. 2018)	ResNet-50-FPN	45.3
iHOI (Xu et al. 2020)	ResNet-50-FPN	45.8
Xu *et al.* (Xu et al. 2019)	ResNet-50-FPN	45.9
No-Frills° (Gupta et al. 2019)	ResNet-152	46.7
Wang *et al.* (Wang et al. 2019a)	ResNet-50-FPN	47.3
RPNN† (Zhou and Chi 2019)	ResNet-50 (Mask R-CNN)	47.5
TIN (RP_DC_D) (Li et al. 2019b)	ResNet-50-FPN	47.8
C-HOI (Zhou et al. 2020)	ResNet-50	48.3
IP-Net (Wang et al. 2020)	ResNet-50-FPN	51.0
VSGNet (Ulutan et al. 2020)	NASNet (Zoph et al. 2018)	51.7
PMFNet† (Wan et al. 2019)	ResNet-50-FPN	52.0
Our baseline	ResNet-50-FPN	48.5
PD-Net	ResNet-50-FPN	**52.3**
PD-Net†	ResNet-50-FPN	**53.3**
Our baseline	ResNet-152	48.2
PD-Net	ResNet-152	**52.2**

Table 9 Per verb class AP(%) comparisons between PMFNet and PD-Net† on V-COCO.

Verbs	PMFNet (Wan et al. 2019)	PD-Net†
hold-obj	44.01	**45.07**
sit-instr	29.51	**31.86**
ride-instr	70.33	**71.80**
look-obj	45.22	**46.72**
hit-instr	76.30	**78.57**
hit-obj	**52.28**	50.28
eat-obj	44.55	**47.41**
eat-instr	5.93	**6.94**
jump-instr	**53.39**	52.75
lay-instr	26.40	**28.25**
talk on phone-instr	54.69	**56.64**
carry-obj	44.24	**45.64**
throw-obj	**49.76**	47.82
catch-obj	54.11	**55.01**
cut-instr	40.08	**42.69**
cut-obj	**40.01**	39.24
work on computer-instr	67.39	**67.98**
ski-instr	**53.04**	52.59
surf-instr	80.47	**80.95**
skateboard-instr	86.81	**88.00**
drink-instr	46.76	**53.84**
kick-obj	72.70	**74.50**
read-obj	36.80	**39.07**
snowboard-instr	74.33	**76.55**
mean	52.05	**53.34**

is chosen as it is very similar to our baseline. Results show that PD-Net achieves superior performance on all of these top 10 verbs.

5.3.2 Performance comparisons on V-COCO

To boost the performance on V-COCO, we add another appearance feature stream to both our baseline and PD-Net, following (Wan et al. 2019). There are consequently a total of five feature streams for experiments on V-COCO. This new stream extracts appearance features from union boxes composed of human-object pairs. We further apply LPCA to this feature stream in PD-Net. As shown in Table 8, PD-Net outperforms state-of-the-art methods by clear margins with both object detectors. In particular, PD-Net outperforms one of the most recently developed methods, i.e. VSGNet (Ulutan et al. 2020). As shown in Table 2, the object detector utilized by VSGNet is much stronger (Huang et al. 2017) than ours; nevertheless, PD-Net still outperforms VSGNet by clear margins, as indicated in Table 8.

Table 10 Performance Comparisons on HOI-VP.

Method	Feature Extraction Backbone	mAP
iCAN (Gao et al. 2018)	ResNet-50	58.32
TIN (Li et al. 2019b)	ResNet-50	60.66
No-Frills (Gupta et al. 2019)	ResNet-152	61.05
Peyre *et al.* (Peyre et al. 2019)	ResNet-50-FPN	61.46
PMFNet (Wan et al. 2019)	ResNet-50-FPN	62.30
Our baseline	ResNet-50	61.18
PD-Net	ResNet-50	**63.11**
Our baseline	ResNet-50-FPN	61.10
PD-Net	ResNet-50-FPN	**63.66**
Our baseline	ResNet-152	60.69
PD-Net	ResNet-152	**63.60**

Moreover, PD-Net outperforms another particularly strong model, named PMFNet (Wan et al. 2019) by 0.3% (52.3%-52.0%) in mAP. The excellent performance of PMFNet may benefit from the use of human part features. Therefore, we adopt the same five feature streams that include the human part features in PMFNet as input for PD-Net; this model is denoted as PD-Net† in Table 8. The contributions in this paper remain unchanged. PD-Net† outperforms PMFNet by a large margin of 1.3% (53.3%-52.0%) in mAP. Moreover, as shown in Table 9, we compare the performance between PD-Net† and PMFNet on each of the 24 verbs in V-COCO. Here, our method demonstrates superior performance on the vast majority of verb classes.

5.3.3 Performance comparisons on HOI-VP

We next compare the performance of PD-Net with some recent open-source methods, i.e. iCAN (Gao et al. 2018), TIN (Li et al. 2019b), No-Frills (Gupta et al. 2019), and PMFNet (Wan et al. 2019) on the new HOI-VP database. We also reproduce the method presented in (Peyre et al. 2019) that achieves high performance on the HICO-DET database. To facilitate fair comparison, we compare the performance of PD-Net with each of these methods using the same feature extraction backbone, respectively. As shown in Table 10, PD-Net consistently achieves the best performance out of all compared methods. In particular, PD-Net outperforms one recent powerful model PMFNet by a clear margin of 1.36% (63.66%-62.30%) in mAP. As the verbs (predicates) in the HOI-VP database are very common and polysemic in real world scenarios, experimental results on this database demonstrate the superiority of PD-Net to overcome the verb polysemy problem.

5.4 Qualitative Visualization Results

Fig. 6 illustrates attention scores produced by PAMF for four types of features. HOI categories in this figure share the verb "ride", but differ dramatically in semantic meanings. The "person" proposal in Fig. 6(a) is very small and severely

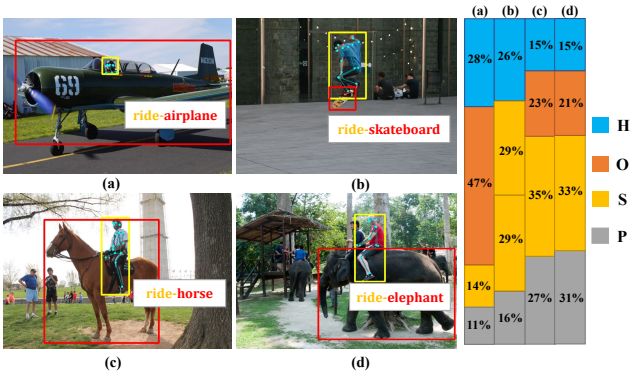

Fig. 6 Attention scores produced by PAMF on four types of features. HOI categories in this figure share the same verb, i.e. "ride". **H**, **O**, **S**, and **P** denote human appearance, object appearance, spatial feature, and human pose feature respectively.

occluded while the "airplane" proposal is very large; therefore, object appearance feature is much more important for verb classification than the human appearance feature. In Fig. 6(b), both the spatial feature and the object appearance feature play important roles in determining the verb. Attention scores for Fig. 6(c) and (d) are similar, as $<person\ ride\ horse>$ and $<person\ ride\ elephant>$ are indeed close in semantics.

Fig. 7, Fig. 8, and Fig. 9 provide more examples that demonstrate PD-Net's advantages in deciphering the verb polysemy problem on HICO-DET, V-COCO, and HOI-VP, respectively. The performance gain by PD-Net compared with our baseline reaches 10.6%, 3.33%, and 48.1% in AP for the "open microwave", "carry backpack", and "play drum" category on the three datasets, respectively.

6 Conclusion

The verb polysemy problem is relatively underexplored and is sometimes even ignored in existing works for HOI detection. Accordingly, in this paper, we propose a novel model named PD-Net, which significantly mitigates the challenging verb polysemy problem. PD-Net includes four novel components: LPCA, LPFA, PAMF, and CSP classifiers. LPCA and LPFA are introduced to generate polysemy-aware visual features. PAMF highlights important feature types for each HOI category. The CSP verb classifiers not only relieve the verb polysemy problem, but also is capable of handling the zero- or few-shot learning problems. Exhaustive ablation studies are performed to demonstrate the effectiveness of these components. We further develop and present a new dataset, named HOI-VP, that is specifically designed to expedite the research on the verb polysemy problem for HOI detection. Finally, by decoding the verb polysemy, we achieve state-of-the-art methods on the three HOI detection benchmarks.

References

Chao YW, Liu Y, Liu X, Zeng H, Deng J (2018) Learning to detect human-object interactions. In: Proc. IEEE Winter Conf. Appl. Comput. Vis., pp 381–389.

Chen L, Zhang H, Xiao J, Nie L, Shao J, Liu W, Chua TS (2017) Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit., pp 5659–5667

Chen X, Gupta A (2017) An implementation of faster rcnn with study for region sampling. arXiv preprint arXiv:170202138

Choi W, Chao YW, Pantofaru C, Savarese S (2015) Indoor Scene Understanding with Geometric and Semantic Contexts. Int J Comput Vis 112(2):204–220

Fang HS, Xie S, Tai YW, Lu C (2017) Rmpe: Regional multi-person pose estimation. In: Proc. IEEE Int. Conf. Comput. Vis., pp 2334–2343

Gao C, Zou Y, Huang JB (2018) ican: Instance-centric attention network for human-object interaction detection. In: Proc. Br. Mach. Vis. Conf., p 41

Gao P, Jiang Z, You H, Lu P, Hoi SC, Wang X, Li H (2019) Dynamic fusion with intra-and inter-modality attention flow for visual question answering. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit., pp 6639–6648

Girdhar R, Ramanan D (2017) Attentional pooling for action recognition. In: Proc. Adv. Neural Inf. Process. Syst., pp 34–45

Girshick R, Radosavovic I, Gkioxari G, Dollár P, He K (2018) Detectron. https://github.com/facebookresearch/detectron

Gkioxari G, Girshick R, Dollár P, He K (2018) Detecting and recognizing human-object interactions. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit., pp 8359–8367

Gu J, Zhao H, Lin Z, Li S, Cai J, Ling M (2019) Scene graph generation with external knowledge and image reconstruction. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit., pp 1969–1978

Gupta S, Malik J (2015) Visual semantic role labeling. arXiv preprint arXiv:150504474

Gupta T, Schwing A, Hoiem D (2019) No-frills human-object interaction detection: Factorization, layout encodings, and training techniques. In: Proc. IEEE Int. Conf. Comput. Vis., pp 9677–9685

He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit., pp 770–778

He K, Gkioxari G, Dollár P, Girshick R (2017) Mask r-cnn. In: Proc. IEEE Int. Conf. Comput. Vis., pp 2961–2969

Hu J, Shen L, Sun G (2018) Squeeze-and-excitation networks. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit., pp 7132–7141

Huang J, Rathod V, Sun C, Zhu M, Korattikara A, Fathi A, Fischer I, Wojna Z, Song Y, Guadarrama S, et al. (2017) Speed/accuracy trade-offs for modern convolutional object detectors. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit., pp 7310–7311

Jaderberg M, Simonyan K, Zisserman A, et al. (2015) Spatial transformer networks. In: Proc. Adv. Neural Inf. Process. Syst., pp 2017–2025

Ji J, Krishna R, Fei-Fei L, Niebles JC (2020) Action genome: Actions as compositions of spatio-temporal scene graphs. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit., pp 10236–10247

Kingma DP, Ba J (2014) Adam: A method for stochastic optimization. arXiv preprint arXiv:14126980

Krishna R, Zhu Y, Groth O, Johnson J, Hata K, Kravitz J, Chen S, Kalantidis Y, Li LJ, Shamma DA, et al. (2017) Visual genome: Connecting language and vision using crowdsourced dense image annotations. Int J Comput Vis (1):32–73

Kuznetsova A, Rom H, Alldrin N, Uijlings J, Ferrari V (2020) The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. Int J Comput

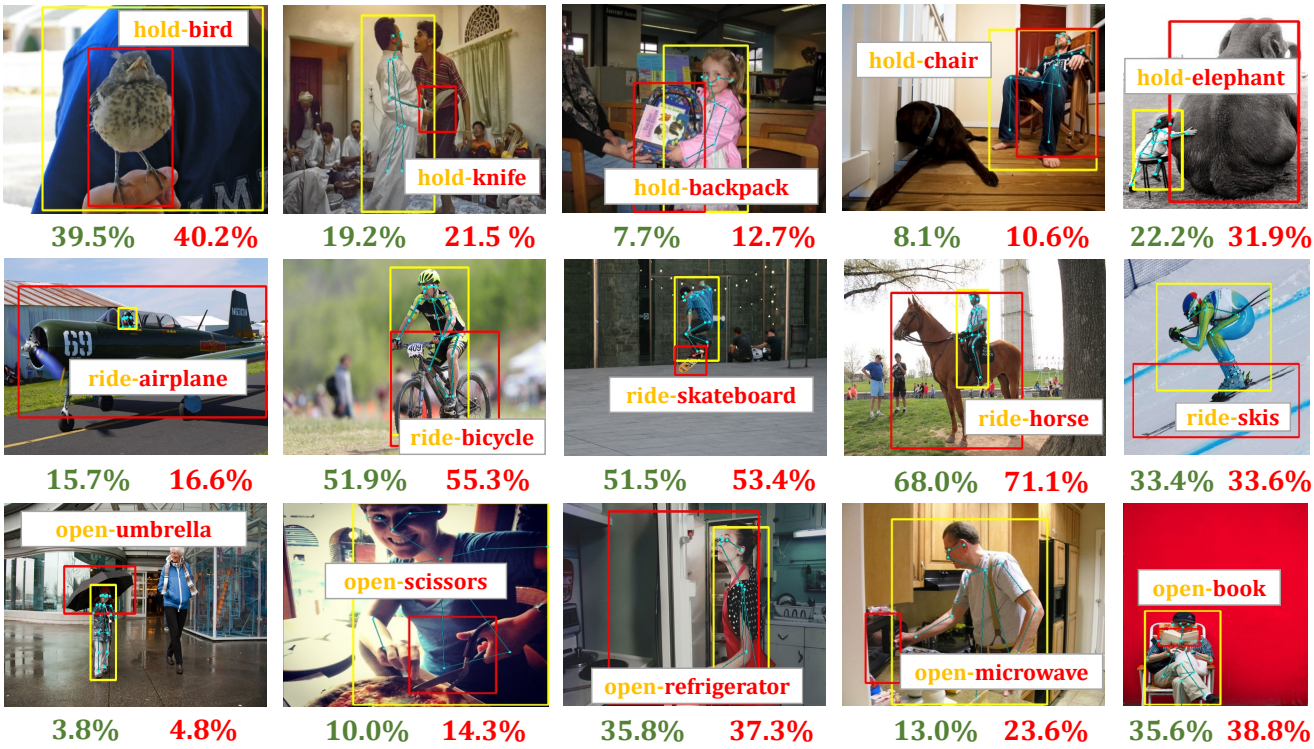

Fig. 7 Visualization of PD-Net's advantage in deciphering the verb polysemy problem on HICO-DET. We randomly select three verbs affected by the polysemy problem: "hold" (top row), "ride" (middle row), and "open" (bottom row). The green and red numbers denote the AP of our baseline and PD-Net respectively for the same HOI category.

Vis 128(7):1956–1981

Li B, Liang J, Wang Y (2019a) Compression artifact removal with stacked multi-context channel-wise attention network. In: Proc. IEEE Int. Conf. Image Process., pp 3601–3605

Li W, Zhu X, Gong S (2018) Harmonious attention network for person re-identification. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit., pp 2285–2294

Li YL, Zhou S, Huang X, Xu L, Ma Z, Fang HS, Wang Y, Lu C (2019b) Transferable interactiveness knowledge for human-object interaction detection. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit., pp 3585–3594

Li YL, Liu X, Lu H, Wang S, Liu J, Li J, Lu C (2020a) Detailed 2d-3d joint representation for human-object interaction. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit., pp 10166–10175

Li YL, Xu L, Liu X, Huang X, Xu Y, Wang S, Fang HS, Ma Z, Chen M, Lu C (2020b) Pastanet: Toward human activity knowledge engine. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit., pp 382–391

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

Lin TY, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollár P, Zitnick CL (2014) Microsoft coco: Common objects in context. In: Proc. Eur. Conf. Comput. Vis., pp 740–755

Lin TY, Dollár P, Girshick R, He K, Hariharan B, Belongie S (2017) Feature pyramid networks for object detection. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit., pp 2117–2125

Lin X, Ding C, Zeng J, Tao D (2020) Gps-net: Graph property sensing network for scene graph generation. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit., pp 3746–3753

Lu C, Krishna R, Bernstein M, Fei-Fei L (2016) Visual relationship detection with language priors. In: Proc. Eur. Conf. Comput. Vis., pp 852–869

MacQueen J, et al. (1967) Some methods for classification and analysis of multivariate observations. In: Proc. the fifth Berkeley symposium on mathematical statistics and probability, vol 1, pp 281–297

Marino K, Rastegari M, Farhadi A, Mottaghi R (2019) Ok-vqa: A visual question answering benchmark requiring external knowledge. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit., pp 3195–3204

Massa F, Girshick R (2018) maskrcnn-benchmark: Fast, modular reference implementation of Instance Segmentation and Object Detection algorithms in PyTorch. https://github.com/facebookresearch/maskrcnn-benchmark

Meng L, Zhao B, Chang B, Huang G, Sun W, Tung F, Sigal L (2019) Interpretable spatio-temporal attention for video action recognition. In: Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCVW), pp 1513–1522

Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J (2013) Distributed representations of words and phrases and their compositionality. In: Proc. Adv. Neural Inf. Process. Syst., pp 3111–3119

Newell A, Yang K, Deng J (2016) Stacked hourglass networks for human pose estimation. In: Proc. Eur. Conf. Comput. Vis., pp 483–499

Pereira S, Pinto A, Amorim J, Ribeiro A, Alves V, Silva CA (2019) Adaptive feature recombination and recalibration for semantic segmentation with fully convolutional networks. IEEE Trans Med Imag 38(12):2914–2925

Peyre J, Laptev I, Schmid C, Sivic J (2019) Detecting unseen visual relations using analogies. In: Proc. IEEE Int. Conf. Comput. Vis., pp 1981–1990

Qi S, Wang W, Jia B, Shen J, Zhu SC (2018) Learning human-object interactions by graph parsing neural networks. In: Proc. Eur. Conf. Comput. Vis., pp 401–417

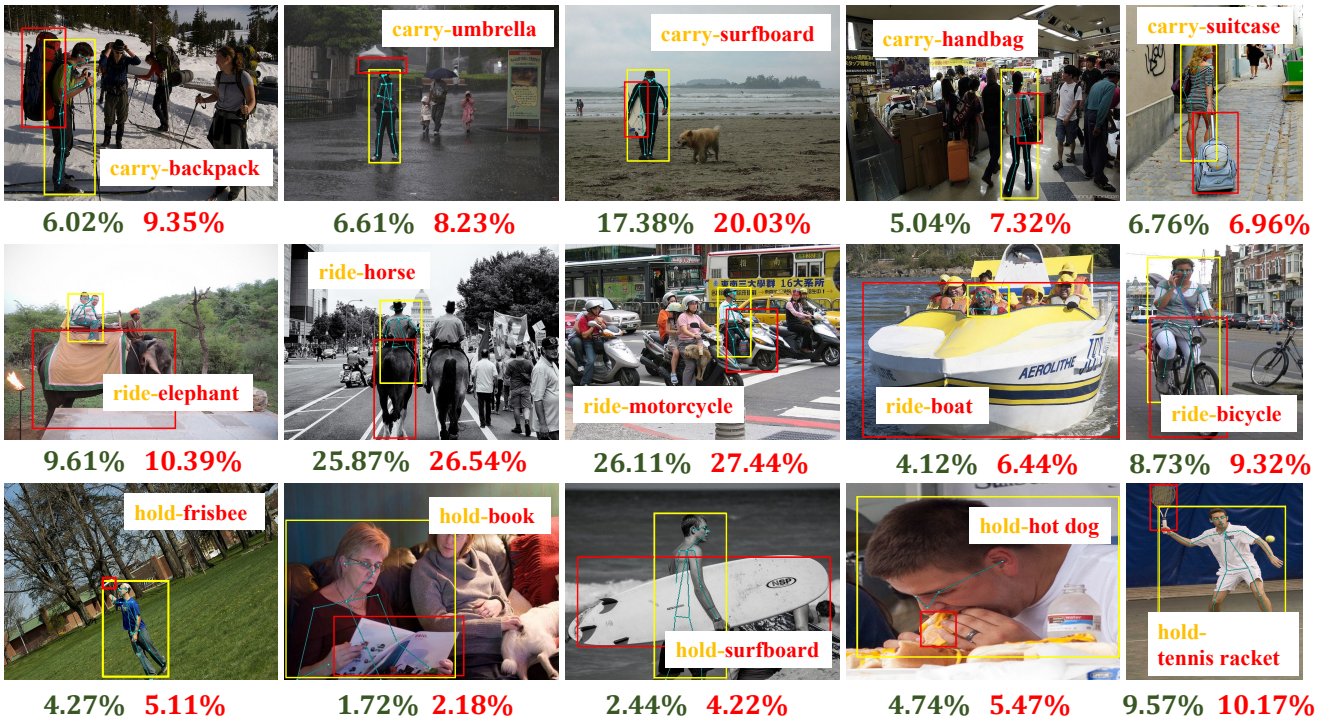

Fig. 8 Visualization of PD-Net's advantage in deciphering the verb polysemy problem on V-COCO. We randomly select three verbs affected by the polysemy problem: "carry" (top row), "ride" (middle row), and "hold" (bottom row).

Ren S, He K, Girshick R, Sun J (2015) Faster r-cnn: Towards real-time object detection with region proposal networks. In: Proc. Adv. Neural Inf. Process. Syst., pp 91–99

Shen L, Yeung S, Hoffman J, Mori G, Li FF (2018) Scaling human-object interaction recognition through zero-shot learning. In: Proc. IEEE Winter Conf. Appl. Comput. Vis., pp 1568–1576

Ulutan O, Iftekhar A, Manjunath B (2020) Vsgnet: Spatial attention network for detecting human object interactions using graph convolutions. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit., pp 13617–13626

Wan B, Zhou D, Liu Y, Li R, He X (2019) Pose-aware multi-level feature network for human object interaction detection. In: Proc. IEEE Int. Conf. Comput. Vis., pp 9469–9478

Wang F, Jiang M, Qian C, Yang S, Li C, Zhang H, Wang X, Tang X (2017) Residual attention network for image classification. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit., pp 3156–3164

Wang T, Anwer RM, Khan MH, Khan FS, Pang Y, Shao L, Laaksonen J (2019a) Deep contextual attention for human-object interaction detection. In: Proc. IEEE Int. Conf. Comput. Vis., pp 5694–5702

Wang T, Yang T, Danelljan M, Khan FS, Zhang X, Sun J (2020) Learning human-object interaction detection using interaction points. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit., pp 4116–4125

Wang W, Wang R, Shan S, Chen X (2019b) Exploring context and visual pattern of relationship for scene graph generation. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit., pp 8188–8197

Xu B, Wong Y, Li J, Zhao Q, Kankanhalli MS (2019) Learning to detect human-object interactions with knowledge. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit., pp 2019–2028

Xu B, Li J, Wong Y, Zhao Q, Kankanhalli MS (2020) Interact as you intend: Intention-driven human-object interaction detection. IEEE Trans Multimedia 22(6):1423–1432

Xu K, Ba J, Kiros R, Cho K, Courville A, Salakhudinov R, Zemel R, Bengio Y (2015) Show, attend and tell: Neural image caption generation with visual attention. In: Proc. Int. Conf. Mach. Learn., pp 2048–2057

Yao T, Pan Y, Li Y, Mei T (2019) Hierarchy parsing for image captioning. arXiv preprint arXiv:190903918

Ye Q, Yuan S, Kim TK (2016) Spatial attention deep net with partial pso for hierarchical hybrid hand pose estimation. In: Proc. Eur. Conf. Comput. Vis., pp 346–361

You Q, Jin H, Wang Z, Fang C, Luo J (2016) Image captioning with semantic attention. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit., pp 4651–4659

Zellers R, Yatskar M, Thomson S, Choi Y (2018) Neural motifs: Scene graph parsing with global context. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit., pp 5831–5840

Zhang H, Kyaw Z, Chang SF, Chua TS (2017) Visual translation embedding network for visual relation detection. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit., pp 5532–5540

Zhao Y, Xiong Y, Wang L, Wu Z, Tang X, Lin D (2020) Temporal Action Detection with Structured Segment Networks. Int J Comput Vis 128(1):74–95

Zheng B, Zhao Y, Yu J, Ikeuchi K, Zhu SC (2015) Scene Understanding by Reasoning Stability and Safety. Int J Comput Vis 112(2):221–238

Zhong X, Ding C, Qu X, Tao D (2020) Polysemy deciphering network for human-object interaction detection. In: Proc. Eur. Conf. Comput. Vis.

Zhou L, Palangi H, Zhang L, Hu H, Corso JJ, Gao J (2019a) Unified vision-language pre-training for image captioning and vqa. arXiv preprint arXiv:190911059

Zhou P, Chi M (2019) Relation parsing neural network for human-object interaction detection. In: Proc. IEEE Int. Conf. Comput. Vis., pp 843–851

Zhou T, Wang W, Qi S, Ling H, Shen J (2020) Cascaded human-object interaction recognition. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit., pp 4263–4272

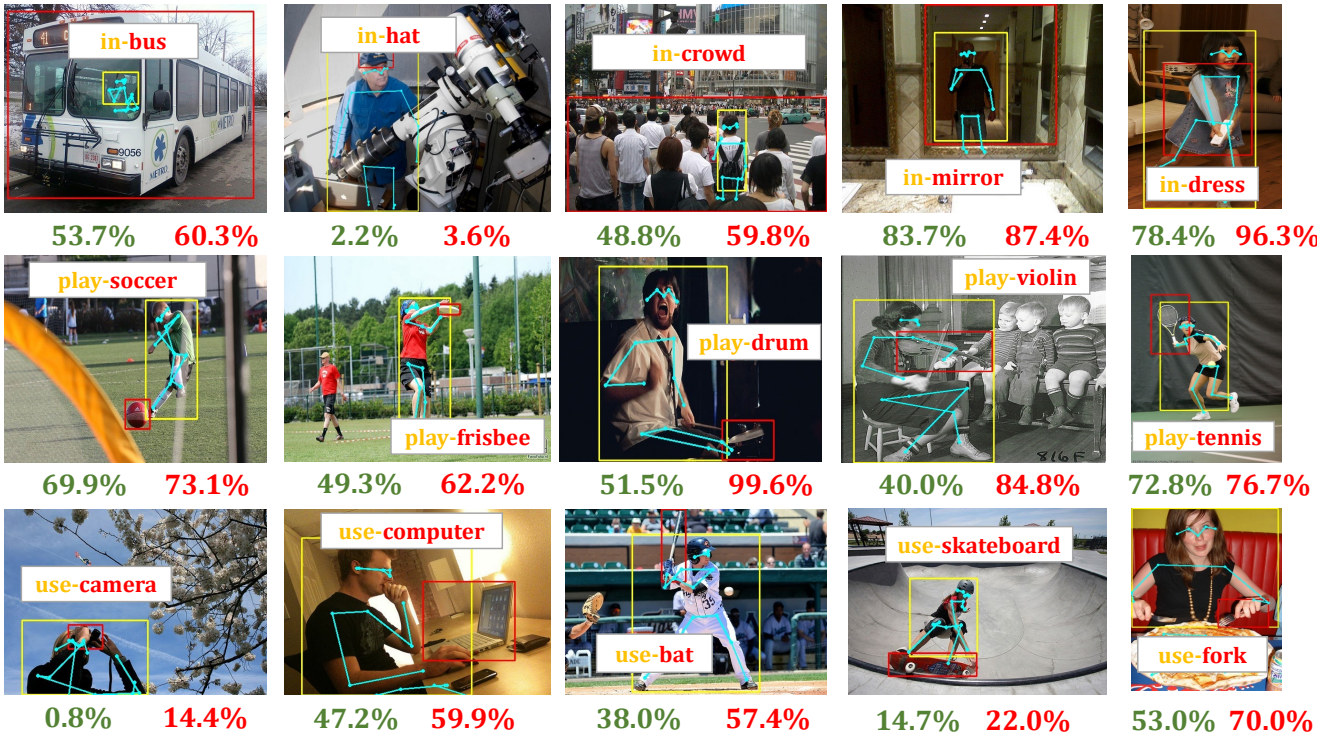

Fig. 9 Visualization of PD-Net's advantage in deciphering the verb polysemy problem on HOI-VP. We randomly select three verbs affected by the polysemy problem: "in" (top row), "play" (middle row), and "use" (bottom row).

Zhou X, Wang D, Krähenbühl P (2019b) Objects as points. In: arXiv preprint arXiv:1904.07850

Zhu Y, Zhao C, Guo H, Wang J, Zhao X, Lu H (2018) Attention couplenet: Fully convolutional attention coupling network for object detection. IEEE Trans Image Process 28(1):113–126

Zhuang B, Wu Q, Shen C, Reid I, Hengel Avd (2017) Care about you: towards large-scale human-centric visual relationship detection. arXiv preprint arXiv:170509892

Zoph B, Vasudevan V, Shlens J, Le QV (2018) Learning transferable architectures for scalable image recognition. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit., pp 8697–8710