Polysemy Deciphering Network for Human-Object Interaction Detection

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

[1] School of Electronic and Information Engineering, South China University of Technology
[2] UBTECH Sydney AI Centre, School of Computer Science, Faculty of Engineering, The University of Sydney, Darlington, NSW 2008, Australia
{eexubin, eequxian971017}@mail.scut.edu.cn, chxding@scut.edu.cn, dacheng.tao@sydney.edu.au

Abstract. Human-Object Interaction (HOI) detection is important in human-centric scene understanding. Existing works typically assume that the same verb in different HOI categories has similar visual characteristics, while ignoring the diverse semantic meanings of the verb. To address this issue, in this paper, we propose a novel Polysemy Deciphering Network (PD-Net), which decodes the visual polysemy of verbs for HOI detection in three ways. First, PD-Net augments human pose and spatial features for HOI detection using language priors, enabling the verb classifiers to receive language hints that reduce the intra-class variation of the same verb. Second, we introduce a novel Polysemy Attention Module (PAM) that guides PD-Net to make decisions based on more important feature types according to the language priors. Finally, the above two strategies are applied to two types of classifiers for verb recognition, i.e., object-shared and object-specific verb classifiers, whose combination further relieves the verb polysemy problem. By deciphering the visual polysemy of verbs, we achieve the best performance on both HICO-DET and V-COCO datasets. In particular, PD-Net outperforms state-of-the-art approaches by 3.81% mAP in the Known-Object evaluation mode of HICO-DET. Code of PD-Net is available at `https://github.com/MuchHair/PD-Net`.

Keywords: Human-Object Interaction, Verb Polysemy, Attention Model

1 Introduction

In recent years, researchers working in the field of computer vision have paid increasing attention to scene understanding tasks [1], [2], [5], [12], [18], [28]. As human beings are often central to real-world scenes, Human-Object Interaction (HOI) detection has become a fundamental problem in scene understanding. HOI detection not only involves 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

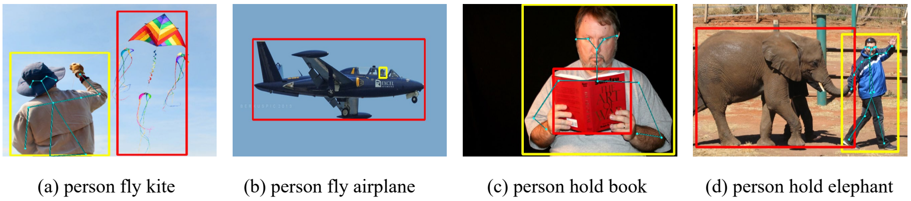

Fig. 1. Examples reflecting the polysemy problem of the same verb. (a) and (b) illustrate HOI examples of "fly". (c) and (d) present HOI examples of "hold".

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 sit on airplane*> (see Fig. 1(b)). Therefore, HOI detection requires multi-label verb classification for each human-object pair.

The HOI detection task is, however, challenging [4], [9]. One major reason is that verbs can be polysemic. As illustrated in Fig. 1, a verb may present substantially different semantic meanings and visual characteristics with respect to different objects. This semantic difference can be very large, resulting in the importance of the same type of visual feature varies dramatically as the objects of interest change. For example, the human pose plays a vital role in describing <*person fly kite*> in Fig. 1(a). However, human pose is invisible and therefore useless for characterizing <*person fly airplane*> in Fig. 1(b). Another example can be found in Fig. 1(c) and (d), here both <*person hold book*> and <*person hold elephant*> pay attention to the spatial feature, i.e. the relative location between two bounding boxes, however, their spatial features are extremely different. Therefore, verb polysemy is a significant challenge in HOI detection.

The verb polysemy problem is relatively underexplored and sometimes even ignored in existing works [9], [14], [16], [27], [29]. Most contemporary approaches assume that the same verb in different HOI categories have similar visual characteristics, so design object-shared verb classifiers. However, due to the polysemic nature of the verbs, a dramatic semantic gap may exist for the same verb across different HOI categories. Chao *et al.* [4] constructed object-specific verb classifiers for each HOI category, which can overcome the polysemy problem for HOI categories with large training data. However, this approach lacks few- and zero-shot learning abilities for HOI categories with small training data.

Accordingly, in this paper, we propose a novel Polysemy Deciphering Network (PD-Net) that addresses the verb polysemy problem in three ways. First, 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, where each binary classifier is used for one verb category identification, respectively. The binary classifiers share parameters expect for the final layers for binary prediction. Therefore, their total model size is equal to one common multi-label verb classifier. The main difference between the binary classifiers lies in their input human pose and spatial features. More specifically, we augment the two features of each

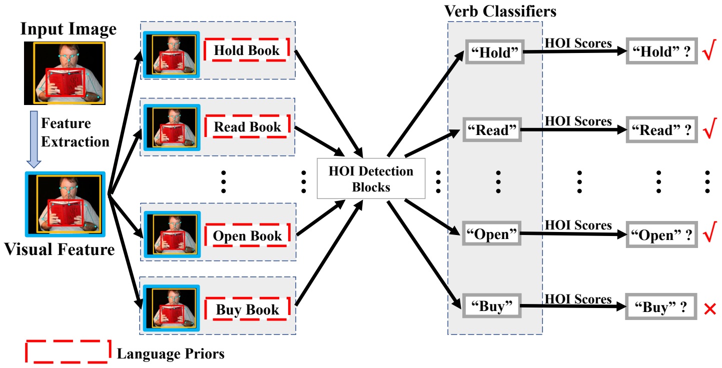

Fig. 2. Visual features of each human-object pair are duplicated multiple times so that the pose and spatial features can be augmented using different language priors. Under the guidance of language priors, polysemy-sensing visual feature of the same verbs can be generated. To reduce the number of duplicated human-object pairs, meaningless HOI categories are ignored, e.g. <*person eat book*> and <*person ride book*>. Meaningful and common HOI categories are available in each popular HOI detection database.

human-object pair with language priors, respectively. These language priors are word embeddings of phrases composed of one verb and one object. The object class is predicted by the Faster R-CNN [24] backbone model; the verb is the one to be determined by one specific binary verb classifier. Our motivation is that both pose and spatial features are vague and varied dramatically for the same verb, as illustrated in Fig. 1. By concatenating the language priors, the classifiers receive hints to reduce the intra-class variation of the same verb for the pose and spatial features.

Moreover, we design a novel Polysemy Attention Module (PAM) that produces attention scores based on the above language priors to dynamically fuse four feature types for each binary classifier: human appearance, object appearance, human pose, and spatial features. The language priors provide hints as to the importance of the features for each HOI category. For example, human pose feature is discriminative when the language prior is "fly kite" (Fig. 1(a)); but is less useful when the language prior is "fly airplane" (Fig. 1(b)). Therefore, PAM deciphers the verb polysemy problem by highlighting more important features for each HOI category.

The above two strategies can be applied to both object-shared and object-specific verb classifiers. We further promote the performance of PD-Net by fusing the two types of classifiers. The classifiers are complementary, as the first can reduce the class imbalance problem for HOI categories with a small sample size, while the second is better able to capture the semantic gap of the same verb in different HOI categories when training data are sufficient. To reduce model size, some of their parameters are shared. In the training stage, the two types of classifier are trained simultaneously, and during testing, their predictions are fused by multiplication. We further propose a clustering-based method to reduce the number of object-specific verb classifiers and handle their few- and zero-shot learning problems.

To the best of our knowledge, PD-Net is the first approach that explicitly handles the verb polysemy problem in HOI detection. We demonstrate the effectiveness of PD-Net on the two most popular benchmarks, i.e. HICO-DET [4] and VCOCO [13]. On both databases, PD-Net achieves significantly better performance than state-of-the-art works.

2 Related Works

Human-Object Interaction Detection. HOI detection performs multi-label verb classification for each human-object pair, meaning that multiple verbs may be used to describe the interaction between the same human-object pair. Depending on the order of verb classification and target object association, existing approaches in HOI detection can be divided into two main sets of methods. The first set of methods infer the verbs performed by one person, then associate each verb of the person with the target objects in the image. Multiple target object association approaches have been proposed. For example, Shen *et al.* [25] proposed an approach based on the value of object detection scores. Gkioxari *et al.* [11] fit a distribution density function of the target object locations based on the human appearance feature. Qi *et al.* [23] adopted a graph parsing network to associate the target objects. Zhou *et al.* [35] constructed graph models that highlight the importance of the link between object and human body parts.

The second category of methods first pair each human instance with all object instances as candidates, then recognize the verb for each candidate pair [14]. For example, Xu *et al.* [32] constructed a graph neural network for verb classification. Xu *et al.* [31] utilized the gaze and intention of human to assist HOI detection. Li *et al.* [16] introduced a Transferable Interactiveness Network to exclude candidate pairs without interactions. Wan *et al.* [27] employed human pose as cues [6], [7] to obtain important human part features for verb classification. Wang *et al.* [29] extracted context-aware human and object appearance features in order to promote HOI detection performance. Peyre *et al.* [22] constructed a multi-stream model that projects the visual features and word embeddings into a joint space, which is powerful for detecting unseen HOI categories.

The Exploitation of Language Priors. Language priors have been successfully utilized in many computer vision tasks, including Scene Graph Generation [18], [19], [30], Image Captioning [3], [15], [33], [34] and Visual Question Answering [10], [20], [26]. Moreover, there are several works that have also adopted language priors for HOI detection [22], [32]. They improve HOI detection by exploiting the correlation between similar verbs or HOI categories with the help of lanuage priors. However, these works do not employ language priors to solve the verb polysemy problem.

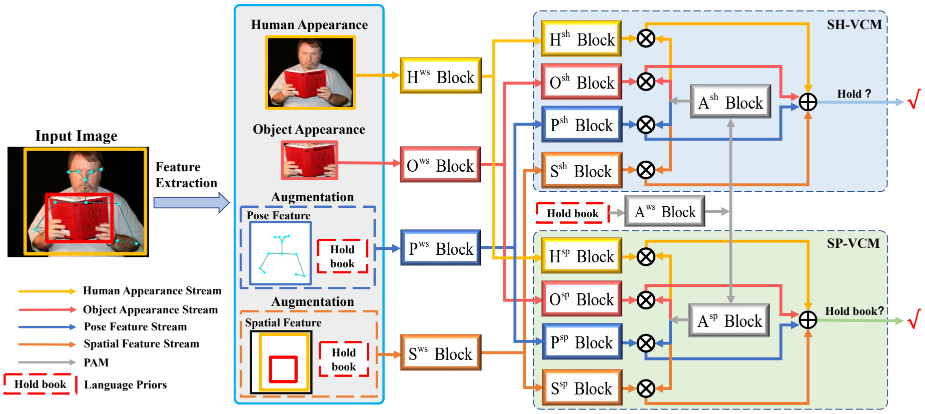

Fig. 3. Overview of the Polysemy Deciphering Network. In the interests of simplicity, only one binary object-shared classifier for ”hold” and one binary object-specific classifier for ”hold book” are illustrated. $\mathbf{A}^{ws}$, $\mathbf{A}^{sh}$, and $\mathbf{A}^{sp}$ are blocks for the PAM module. $\mathbf{H}^{ws}$, $\mathbf{O}^{ws}$, $\mathbf{P}^{ws}$, and $\mathbf{S}^{ws}$ are blocks shared by SH-VCM and SP-VCM. The superscript ”ws” denotes ”weight sharing”. $\mathbf{H}^{sh}$ ($\mathbf{H}^{sp}$), $\mathbf{O}^{sh}$ ($\mathbf{O}^{sp}$), $\mathbf{P}^{sh}$ ($\mathbf{P}^{sp}$) and $\mathbf{S}^{sh}$ ($\mathbf{S}^{sp}$) are blocks in SH-VCM (SP-VCM). The binary classification scores from $\mathbf{H}^{sh}$ ($\mathbf{H}^{sp}$), $\mathbf{O}^{sh}$ ($\mathbf{O}^{sp}$), $\mathbf{P}^{sh}$ ($\mathbf{P}^{sp}$) and $\mathbf{S}^{sh}$ ($\mathbf{S}^{sp}$) are fused by attention scores produced by PAM. $\otimes$ and $\oplus$ denote the element-wise multiplication and addition operations, respectively.

3 Our Method

3.1 Overview

Given an image, we obtain human and object proposals using Faster R-CNN [24]. Each human proposal h and object proposal o will be paired as a candidate for verb classification. PD-Net produces a set of verb classification scores for each candidate pair. As shown in Fig. 2, we first transform the multi-label verb classification into a set of binary classification problems. Moreover, as illustrated in Fig. 3, in order to overcome the verb polysemy problem, PD-Net performs HOI detection by combining Language Prior Augmentation (LPA), Polysemy Attention Module (PAM), Object-SHared Verb Classification Module (SH-VCM), and Object-SPecific Verb Classification Module (SP-VCM). SH-VCM and SP-VCM contain a set of object-shared binary classifiers and object-specific binary classifiers, respectively.

The identification score produced by PD-Net for the verb v in the HOI category (h, o, v) can be denoted as $\mathcal{S}^{\mathbf{PD}}_{(h,o,v)}$:

$$\mathcal{S}^{\mathbf{PD}}_{(h,o,v)} = \mathcal{S}^{\mathbf{SH}}_{(h,o,v)} \times \mathcal{S}^{\mathbf{SP}}_{(h,o,v)}, \tag{1}$$

where $S^{\mathbf{SH}}_{(h,o,v)}$ and $S^{\mathbf{SP}}_{(h,o,v)}$ represent the classification scores for the verb v produced by SH-VCM and SP-VCM, respectively.

3.2 Representation and Classification Networks

As shown in Fig. 3, we construct four complementary feature representation streams for each human-object pair: the human appearance stream (the **H** stream), the object appearance stream (the **O** stream), the human-object spatial stream (the **S** stream) and the human pose stream (the **P** stream). We adopt the same Faster R-CNN model as [14] to obtain human and object proposals. For both human and each object category, we select the top 10 proposals according to the detection scores after non-maximum suppression.

In the below, we introduce each of the four feature streams in detail. It is worth noting that unless otherwise specified, the dimension for each FC layer in the following is set as the dimension of its input feature vector in the interests of simplicity.

Human Appearance Stream. Input feature of the **H** stream is a 2048-dimensional vector for each human proposal. We adopt the same way as [14] to obtain this feature vector from Faster R-CNN. This feature vector passes through three blocks for verb classification: $\mathbf{H}^{ws}$, $\mathbf{H}^{sh}$ in SH-VCM, and $\mathbf{H}^{sp}$ in SP-VCM. $\mathbf{H}^{ws}$ is realized by one FC layer, which is followed by one batch normalization (BN) layer and one ReLU layer. $\mathbf{H}^{sh}$ includes only one K_V-dimensional FC layer, where K_V denotes the number of verb categories. Each element of this layer's output denotes the prediction of one binary classifier according to the **H** stream feature. $\mathbf{H}^{sp}$ includes two FC layers; the first of these is followed by one BN layer and one ReLU layer. The dimension of the second one is K_T, which represents the number of meaningful HOI categories provided by each HOI dataset [4], [13]. It is worth noting that K_V and K_T equals to the number of binary classifiers in SH-VCM and SP-VCM, respectively.

Object Appearance Stream. For the **O** stream, the model structure and the way in which features are constructed are the same as for the **H** stream. Feature vectors for the **O** stream pass through three blocks for verb classification: $\mathbf{O}^{ws}$, $\mathbf{O}^{sh}$ in SH-VCM, and $\mathbf{O}^{sp}$ in SP-VCM.

Spatial Feature Stream. Following [14], we encode a 42-dimensional spatial feature vector using the bounding box coordinates of one human-object pair. As illustrated in Fig. 1, spatial features are vague and vary dramatically for the same verb. By transforming the multi-label verb classifier into a set of binary classifiers, the same spatial feature can be augmented using different language priors to generate polysemy-sensing feature of each specific verb, as shown in Fig. 2. We concatenate this spatial feature with a 600-dimensional word embedding of two words. One word denotes the verb to be identified and the other one is the detected object. Word embeddings are generated using the word2vec tool [21]. Features for this stream pass through three blocks: $\mathbf{S}^{ws}$, $\mathbf{S}^{sh}$ in SH-VCM, and $\mathbf{S}^{sp}$ in SP-VCM. $\mathbf{S}^{ws}$ includes two FC layers. $\mathbf{S}^{sh}$ is a K_V-dimensional FC layer. $\mathbf{S}^{sp}$ incorporates two FC layers; the dimension of the last one is K_T.

Human Pose Feature Stream. We use a pose estimation model [8] to obtain the coordinates of 17 keypoints for each human instance. Following [14], the human keypoints and the bounding box coordinates of object proposal are then encoded into a 272-dimensional pose feature vector. To generate polysemy-sensing features, we use the same strategy as the **S** stream to augment the pose feature with word embeddings. Features for this stream pass through three blocks: $\mathbf{P}^{ws}$, $\mathbf{P}^{sh}$ in SH-VCM, and $\mathbf{P}^{sp}$ in SP-VCM. The structure for each of these blocks is the same as its counterpart in the **S** stream.

It is worth noting that LPA is not applied to human and object appearance features. This is because they are redundant and thus require no further augmentation. In comparisons, the pose and spatial features are simply encoded using box or keypoint coordinates. Therefore, they are insufficient in information.

3.3 Polysemy Attention Module

As mentioned 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 identifying the same verb may vary dramatically as the objects change. As shown in Fig. 1, the human appearance and pose features are 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 PAM 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") as in Section 3.2. PAM can be used for both SH-VCM and SP-VCM, and its structure is very simple. As illustrated in Fig. 3, PAM incorporates three blocks: $\mathbf{A}^{ws}$, $\mathbf{A}^{sh}$, and $\mathbf{A}^{sp}$. The $\mathbf{A}^{ws}$ block is shared by SH-VCM and SP-VCM in order to reduce the number of parameters. The $\mathbf{A}^{ws}$ block is composed of two 600-dimensional FC layers. $\mathbf{A}^{sh}$ is for SH-VCM only and includes only one four-dimensional FC layer. $\mathbf{A}^{sp}$ is for SP-VCM only and includes two FC layers, the dimensions of which are 600 and 4, respectively. The output of the two four-dimensional FC layers are processed by the sigmoid activation function and then used as attention scores for SH-VCM and SP-VCM, respectively. In this way, the role of important features with respect to the language prior is highlighted, while that of less useful features is suppressed.

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

$$\mathcal{S}^{\mathbf{SH}}_{(h,o,v)} = \sigma\Big(\sum_{i \in \{\mathbf{H},\mathbf{O},\mathbf{P},\mathbf{S}\}} a^{\mathbf{sh}}_{(i,o,v)} s^{\mathbf{sh}}_{(i,o,v)}\Big), \tag{2}$$

where i denotes one feature stream, while $a^{\mathbf{sh}}_{(i,o,v)}$ is the attention score generated by PAM for the i-th feature stream. $s^{\mathbf{sh}}_{(i,o,v)}$ is the verb prediction score generated by the i-th feature stream. $\sigma(\cdot)$ represents the sigmoid activation function.

3.4 Object-SPecific Verb Classification Module

Since the verb polysemy problem is essentially caused by the change of objects, SP-VCM is introduced to construct a binary classifier for each meaningful HOI category. Meaningful and common HOI categories are provided by each popular HOI detection database, e.g. HICO-DET [4] and V-COCO [13]. Meaningless HOI categories, e.g. <*person read bicycle*> and <*person open bicycle*>, are not considered in HOI detection. The input features for SH-VCM and SP-VCM are identical. The four feature streams for each binary classifier in SP-VCM are also fused using the attention scores produced by PAM. In theory, SP-VCM will be better at solving the polysemy problem than SH-VCM, provided that each HOI category has sufficient training data; however, due to the class imbalance problem, only limited training data is available for many HOI categories. Therefore, we propose to combine SH-VCM and SP-VCM for verb classification purposes. In addition, some of their layers are shared between them, i.e. $\mathbf{H}^{ws}$, $\mathbf{O}^{ws}$, $\mathbf{P}^{ws}$, and $\mathbf{S}^{ws}$, to reduce the overfitting risk of SP-VCM on detecting HOI categories that have limited training data. In the experiments section, we prove that in the absence of this parameter sharing, the naive combination of SH-VCM and SP-VCM achieves worse performance on detecting HOI categories which have limited training data.

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

$$\mathcal{S}^{\mathbf{SP}}_{(h,o,v)} = \sigma(\sum_{i \in \{\mathbf{H},\mathbf{O},\mathbf{P},\mathbf{S}\}} a^{\mathbf{sp}}_{(i,o,v)} s^{\mathbf{sp}}_{(i,o,v)}), \tag{3}$$

where i denotes one feature stream. $a^{\mathbf{sp}}_{(i,o,v)}$ and $s^{\mathbf{sp}}_{(i,o,v)}$ represent the attention score and the verb identification score for the i-th feature stream, respectively.

3.5 Clustering-based SP-VCM

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 is therefore $|O| \times |V|$, which is usually very large. Accordingly, SP-VCM may include many binary classifiers, even if we remove meaningless HOI categories. In the following, we propose a clustering-based method, named CSP-VCM, aiming at reducing the number of binary classifiers in SP-VCM.

The main motivation for this step is that some HOI categories with the same verb are semantically similar [2], [22], e.g. <*person hold elephant*>, <*person hold horse*> and <*person hold cow*>; therefore, they can share the same object-specific classifier. In this way, the number of binary classifiers in CSP-VCM is reduced. In more detail, we first obtain all meaningful and common HOI categories for each verb. The number of meaningful HOI categories that include the verb v is indicated by O_v. Next, we use the K-means method to cluster the HOI categories with the same verb v into C_v clusters according to the cosine distance between the word embeddings of objects. We empirically set C_v of 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-VCM. For example, a new HOI category <*person hold sheep*> during testing can share the same classifier with HOI categories that have similar semantic meanings, e.g. $<$*person hold horse*$>$.

3.6 Training and Testing

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

$$\mathcal{L}_{(h,o,v)} = \mathcal{L}^{\mathbf{SH}}(\mathcal{S}^{\mathbf{SH}}_{(h,o,v)}, l_v) + \mathcal{L}^{\mathbf{SP}}(\mathcal{S}^{\mathbf{SP}}_{(h,o,v)}, l_v), \tag{4}$$

where $\mathcal{L}^{\mathbf{SH}}$ and $\mathcal{L}^{\mathbf{SP}}$ represent the loss of the corresponding binary classifiers in SH-VCM and SP-VCM, respectively. The loss functions of all binary classifiers are realized by binary cross-entropy loss. l_v denotes a binary label ($l_v \in \{0, 1\}$).
Testing. During testing, we use the same method as that in the training stage to obtain language priors: the object category in the prior is predicted by Faster R-CNN (rather than the ground-truth); the verb category in the prior varies for each binary classifier of the verb. Finally, the prediction score for one HOI category (h, v, o) is represented as follows:

$$\mathcal{S}^{\mathbf{HOI}}_{(h,o,v)} = \mathcal{S}_h \times \mathcal{S}_o \times \mathcal{S}^{\mathbf{PD}}_{(h,o,v)} \times \mathcal{S}^{\mathbf{I}}_{(h,o)}, \tag{5}$$

where $\mathcal{S}_h$ and $\mathcal{S}_o$ are the detection scores of human and object proposals, respectively. $\mathcal{S}^{\mathbf{I}}_{(h,o)}$ denotes the detection score generated by a pre-trained Interactiveness Network (INet) [16], which can suppress pairs that contain no interaction. In the experiments section, we demonstrate that INet slightly promotes the performance of PD-Net.

4 Experiments

4.1 Datasets and Metrics

Datasets. HICO-DET [4] and V-COCO [13] are the two most popular benchmarks for HOI detection. HICO-DET is a large-scale dataset, containing 47,776 images in total. Of these, 38,118 images are assigned to the training set, while the remaining 9568 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. V-COCO [13] is a subset of MS-COCO [17], including 2533, 2867, and 4946 images for training, validation and testing, respectively. Each human instance is annotated with binary labels for 26 action categories.
Metrics. According to the official protocols [4], [13], mean average precision (mAP) is used as the evaluation metric for HOI detection. On both HOI datasets, a positive human-object pair must meet the following two requirements: (1) the predicted HOI category must be the same type as the ground truth; (2) 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 modes of mAP 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, as the categories of objects in all images are known, we need only to compute the AP for each HOI category from images containing the interested object; therefore, the KO mode can better reflect the verb classification ability. Finally, for V-COCO, the role mAP [13] (AP_{role}) is used for evaluation.

4.2 Implementation Details

We adopt the same Faster R-CNN model as [14] for object detection and object feature extraction. To facilitate fair comparison with the majority of existing works [14], [11], [23], [16], we fix parameters of the feature extraction backbone on both datasets. The ratio between positive and negative HOI candidate pairs is set to 1:1000, while the output layer dimensions of K_V and K_T are set to 117 (24) and 600 (234), respectively in HICO-DET (V-COCO). For CSP-VCM, K_T is reduced from 600 (234) to 187 (45) in HICO-DET (V-COCO).

We adopt a two-step training strategy. First, we train all blocks in PD-Net for 6 epochs. Second, we fix those parameters that are shared between SH-VCM and SP-VCM, and fine-tune only their exclusive parameters for 2 epochs. Adam is used to optimize PD-Net with a learning rate of 1e-3 (1e-4) on HICO-DET (V-COCO). During testing, we rank the HOI candidate pairs by their detection scores (according to Equation (5)) and calculate mAP for evaluation purposes.

4.3 Ablation Studies

In this subsection, we perform ablation studies to demonstrate the effectiveness of each of the components in PD-Net on HICO-DET [4]. Experimental results are summarized in Table 1 and Table 2.

Language Prior Augmentation. LPA is used to provide hints for the classifier to reduce the intra-class variation of the pose and spatial features by augmenting them with language priors. When LPA is incorporated, SH-VCM is promoted by 0.42% and 0.73% mAP in the DT and KO modes, respectively (in Table 1).

Polysemy Attention Module. PAM is designed to decipher the verb polysemy by assigning larger weights to the important feature types for each HOI category. As shown in Table 1, under the guidance of PAM, the performance of SH-VCM is improved significantly by 1.10% and 1.34% respectively in the two evaluation modes. We further equip SH-VCM with both LPA and PAM, after which the mAP in the DT (KO) mode is promoted from 17.31% (22.40%) to 18.77% (24.04%). Moreover, the combination of LPA and PAM can improve the performance of SP-VCM by 1.04% (0.84%) in the terms of DT (KO) mAP.

Combination of SH-VCM and SP-VCM. We also naively combine the HOI detection scores from the separately trained SH-VCM+LPA+PAM and SP-HCM+LPA+PAM models, meaning that the two models share no parameters except for their backbone. As shown in Table 1, their combination outperforms

Table 1. Ablation studies of LPA, PAM, CSP-VCM, and INet. Full refers to the evaluation with the mAP metric on all 600 HOI categories of the HICO-DET database.

Method	DT (Full)	KO (Full)
SH-VCM	17.31	22.40
SP-VCM	17.45	22.07
SH-VCM + LPA	17.73	23.13
SH-VCM + PAM	18.41	23.74
SH-VCM + LPA + PAM	18.77	24.04
SP-VCM + LPA + PAM	18.49	22.91
CSP-VCM + LPA + PAM	18.88	23.68
SH-VCM + SP-VCM + LPA + PAM	19.69	24.35
SH-VCM + SP-VCM + LPA + PAM + INet	20.20	24.22

Table 2. Ablation studies of PD-Net on Full, Rare, and Non-Rare HOI categories.

	DT Mode		
Method	Full	Rare	Non-Rare
SH-VCM + SP-VCM + LPA + PAM + INet	20.20	14.08	22.03
SH-VCM + CSP-VCM + LPA + PAM + INet	20.22	14.59	21.90
PD-Net	19.99	14.95	21.50

both SH-VCM+LPA+PAM and SP-VCM+LPA+PAM in both the DT and KO modes. These experimental results clearly demonstrate that the two classifiers are complementary and significant to overcoming the verb polysemy problem.

When INet is integrated, the mAP in DT mode is promoted by 0.51%. However, the mAP in the KO mode decreases by 0.13%. This is because INet can suppress candidate pairs that have no interaction, which are typically caused by the incorrect or redundant object proposals. In comparison, KO mode is less affected by object detection errors; therefore, PD-Net can achieve high performance without the assist from INet in this mode. This experiment demonstrates that the strong performance of PD-Net is primarily due to its excellent verb classification ability.

Clustering-based SP-VCM. As shown in Table 2, the combination of CSP-VCM+LPA+PAM and SH-VCM+LPA+PAM also achieves good performance of 20.22% mAP in the DT Full mode. It outperforms the combination of SP-VCM+LPA+PAM and SH-VCM+LPA+PAM on rare HOI categories by 0.51%. Performance promotion on the rare HOI categories is obtained by sharing verb classifiers with semantically similar HOI categories. These experimental results demonstrate that CSP-VCM is effective in reducing the number of object-specific classifiers and handling their few-shot learning problems in SP-VCM.

The Weight Sharing Strategy. PD-Net shares the weights of some blocks between SH-VCM and SP-VCM. As shown in Table 2, on full HOI categories, this strategy is less effective than the naive combination between SH-VCM and SP-VCM. However, with the weight sharing strategy, PD-Net achieves far better performance on rare HOI categories with a margin of 0.87% in terms of mAP. This experiment justifies the weight sharing strategy in PD-Net.

Table 3. Performance comparisons on HICO-DET. ‡ denotes the methods that adopt exactly the same object detector and feature extraction backbone as PD-Net.

	DT Mode			KO Mode		
Method	Full	Rare	Non-Rare	Full	Rare	Non-Rare
Shen *et al.* [25]	6.46	4.24	7.12	-	-	-
HO-RCNN [4]	7.81	5.37	8.54	10.41	8.94	10.85
InteractNet [11]	9.94	7.16	10.77	-	-	-
GPNN [23]	13.11	9.34	14.23	-	-	-
Xu *et al.* [32]	14.70	13.26	15.13	-	-	-
iCAN [9]	14.84	10.45	16.15	16.26	11.33	17.73
Wang *et al.* [29]	16.24	11.16	17.75	17.73	12.78	19.21
No-Frills‡ [14]	17.18	12.45	18.68	-	-	-
TIN [16]	17.22	13.51	18.32	19.38	15.38	20.57
RPNN [35]	17.35	12.78	18.71	-	-	-
PMFNet [27]	17.46	15.65	18.00	20.34	17.47	21.20
Peyre *et al.* [22]	19.40	14.60	20.90	-	-	-
Our baseline (SH-VCM)	17.31	12.09	18.86	22.40	16.48	24.17
PD-Net	**19.99**	14.95	**21.50**	**24.15**	**18.06**	**25.97**
PD-Net†	**20.81**	**15.90**	**22.28**	**24.78**	**18.88**	**26.54**

Table 4. Performance comparisons on V-COCO. ‡ denotes the methods that adopt exactly the same object detector and feature extraction backbone.

Methods	AP_{role}
Gupta *et al.* [13], [11]	31.8
InteractNet [11]	40.0
GPNN [23]	44.0
iCAN [9]	45.3
Xu *et al.* [32]	45.9
Wang *et al.* [29]	47.3
RPNN [35]	47.5
Baseline of PMFNet [27]	48.6
TIN [16]	48.7
PMFNet$^{\circ}$ [27]	50.9
PMFNet‡ [27]	52.0
Our baseline (SH-VCM)	48.2
PD-Net	51.6
PD-Net‡	**52.6**

4.4 Comparisons with State-of-the-art Methods

In this subsection, we compare the performance of PD-Net with state-of-the-art methods. Experimental results are summarized in Table 3 and Table 4.

HICO-DET. PD-Net outperforms all state-of-the-art methods by considerable margins. In particular, with the same Faster R-CNN model, PD-Net outperforms No-Frills [14] by 2.81%, 2.50%, and 2.82% in mAP on the full, rare and non-rare HOI categories of the DT mode respectively. PD-Net outperforms state-of-the-art approaches by 3.81% mAP in the KO mode. We also observe that the performance of PD-Net can be further promoted by fusing another classifier that is based only on the human appearance feature. Similar to [32], [22], the same 600-dimensional word embedding used in LPA and PAM is projected to a new feature space by two FC layers, the dimensions of which are 1024 and 2048 respectively. The inner product between the human appearance feature and that of the word embedding produces the score for one verb binary classifier. During testing, this score is fused with the original score of PD-Net via multiplication for HOI detection. This fused model is denoted as PD-Net†. As shown in Table 3, PD-Net† outperforms state-of-the-art methods by 1.41% in the DT mode.

V-COCO. Following [27], we add an union-box appearance stream on both our baseline and PD-Net. This stream extracts appearance features from the union bounding boxes composed of human-object proposal pairs. As shown in Table 4, PD-Net achieves 51.6% in mAP and outperforms our baseline by a large margin of 3.4%. The best performance on V-COCO was achieved by PMFNet [27] with 52% in mAP, which is slightly better than PD-Net. The following two details may be helpful in the implementation of PMFNet. First, PMFNet fine-tuned the Feature Pyramid Network (FPN) component of its feature extraction backbone on V-COCO. Second, it adopted human part features [27]. Therefore, we further compare PD-Net with PMFNet using two more fair settings in Table 4. First, we fix the FPN component in PMFNet according to the code released by the authors (denoted as PMFNet$^\circ$) and re-train the model. PD-Net outperforms PMFNet$^\circ$ by 0.7%. Second, we adopt the same feature extraction backbone and the five feature streams as described in [27] for PD-Net. The contributions in this paper remain unchanged. This model is denoted as PD-Net‡ in Table 4. It is shown that PD-Net‡ outperforms PMFNet by 0.6%. The above experiments justify the effectiveness of PD-Net.

4.5 Qualitative Visualization Results

Fig. 4 visualizes PD-Net's advantage in deciphering the polysemy problem of verbs on HICO-DET. The performance gain by PD-Net compared with SH-VCM reaches 23.2% in AP for the "open refrigerator" category.

Fig. 5 illustrates attention scores produced by PAM for four types of features on HICO-DET. HOI categories in this figure share the verb "ride", but differ dramatically in semantic meanings. The "person" proposal in Fig. 5(a) is very small and severely occluded while the "airplane" proposal is very large; thus, object appearance feature is much more important for verb classification than the human appearance feature. In Fig. 5(b), both the spatial feature and human pose feature play important roles in determining the verb. Attention scores for Fig. 5(c) and (d) are similar, as <*person ride horse*> and <*person ride elephant*> are indeed close in semantics. More qualitative visualization results on the V-COCO database are provided in the supplementary file.

5 Conclusions

The verb polysemy problem is relatively underexplored and sometimes even ignored in existing works for HOI detection. In this paper, we propose a novel model, named PD-Net, which significantly mitigates the challenging verb polysemy problems for HOI detection through the use of three key components: LPA,

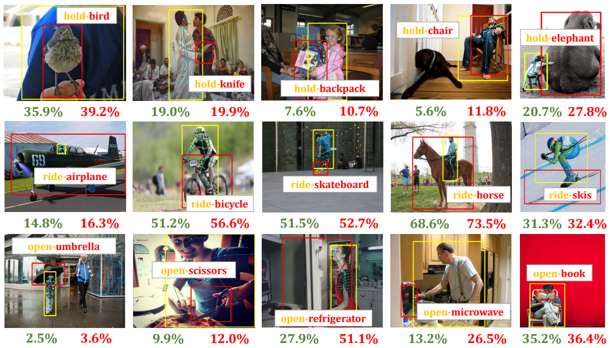

Fig. 4. Visualization of PD-Net's advantage in deciphering the verb polysemy problem. We randomly select three verbs affected by the polysemy problem: "hold" (top row), "ride" (middle row), and "open" (bottom row). The green number and red number denote the AP of SH-VCM and PD-Net respectively for the same HOI category.

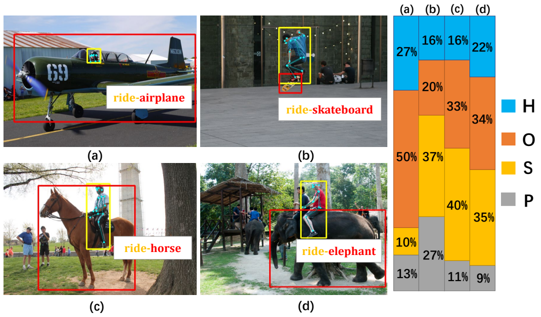

Fig. 5. Attention scores produced by PAM on four types of features. HOI categories in this figure have the same verb ("ride"). **H**, **O**, **S** and **P** denote human appearance, object appearance, spatial feature, and human pose feature respectively.

PAM, and the combination of SH-VCM with SP-VCM. Exhaustive ablation studies are performed to demonstrate the effectiveness of these three components. Finally, the effectiveness of PD-Net for decoding the verb polysemy is demonstrated on the two most popular datasets for HOI detection.

Acknowledgements

Changxing Ding is the corresponding author. This work was supported by the NSF of China under Grant 61702193, the Science and Technology Program of Guangzhou under Grant 201804010272, the Program for Guangdong Introducing Innovative and Entrepreneurial Teams under Grant 2017ZT07X183, the Fundamental Research Funds for the Central Universities of China under Grant 2019JQ01, and ARC FL-170100117.

References

1. Ashual, O., Wolf, L.: Specifying object attributes and relations in interactive scene generation. In: ICCV (2019)
2. Bansal, A., Rambhatla, S.S., Shrivastava, A., Chellappa, R.: Detecting human-object interactions via functional generalization. arXiv preprint arXiv:1904.03181 (2019)
3. Cadene, R., Ben-Younes, H., Cord, M., Thome, N.: Murel: Multimodal relational reasoning for visual question answering. In: CVPR (2019)
4. Chao, Y.W., Liu, Y., Liu, X., Zeng, H., Deng, J.: Learning to detect human-object interactions. In: WACV (2018)
5. Chen, T., Yu, W., Chen, R., Lin, L.: Knowledge-embedded routing network for scene graph generation. In: CVPR (2019)
6. Chen, Y., Wang, Z., Peng, Y., Zhang, Z., Yu, G., Sun, J.: Cascaded pyramid network for multi-person pose estimation. In: CVPR (2018)
7. Fang, H.S., Cao, J., Tai, Y.W., Lu, C.: Pairwise body-part attention for recognizing human-object interactions. In: ECCV (2018)
8. Fang, H.S., Xie, S., Tai, Y.W., Lu, C.: Rmpe: Regional multi-person pose estimation. In: ICCV (2017)
9. Gao, C., Zou, Y., Huang, J.B.: ican: Instance-centric attention network for human-object interaction detection. In: BMVC (2018)
10. Gao, P., Jiang, Z., You, H., Lu, P., Hoi, S.C., Wang, X., Li, H.: Dynamic fusion with intra-and inter-modality attention flow for visual question answering. In: CVPR (2019)
11. Gkioxari, G., Girshick, R., Dollár, P., He, K.: Detecting and recognizing human-object interactions. In: CVPR (2018)
12. Gu, J., Zhao, H., Lin, Z., Li, S., Cai, J., Ling, M.: Scene graph generation with external knowledge and image reconstruction. In: CVPR (2019)
13. Gupta, S., Malik, J.: Visual semantic role labeling. arXiv preprint arXiv:1505.04474 (2015)
14. Gupta, T., Schwing, A., Hoiem, D.: No-frills human-object interaction detection: Factorization, layout encodings, and training techniques. In: ICCV (2019)
15. He, S., Tavakoli, H.R., Borji, A., Pugeault, N.: Human attention in image captioning: Dataset and analysis. In: ICCV (2019)
16. Li, Y.L., Zhou, S., Huang, X., Xu, L., Ma, Z., Fang, H.S., Wang, Y., Lu, C.: Transferable interactiveness knowledge for human-object interaction detection. In: CVPR (2019)
17. 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: ECCV (2014)
18. Lin, X., Ding, C., Zeng, J., Tao, D.: Gps-net: Graph property sensing network for scene graph generation. In: CVPR (2020)
19. Lu, C., Krishna, R., Bernstein, M., Fei-Fei, L.: Visual relationship detection with language priors. In: ECCV (2016)
20. Marino, K., Rastegari, M., Farhadi, A., Mottaghi, R.: Ok-vqa: A visual question answering benchmark requiring external knowledge. In: CVPR (2019)
21. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: NIPS (2013)
22. Peyre, J., Laptev, I., Schmid, C., Sivic, J.: Detecting unseen visual relations using analogies. In: ICCV (2019)

23. Qi, S., Wang, W., Jia, B., Shen, J., Zhu, S.C.: Learning human-object interactions by graph parsing neural networks. In: ECCV (2018)
24. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. In: NIPS (2015)
25. Shen, L., Yeung, S., Hoffman, J., Mori, G., Li, F.F.: Scaling human-object interaction recognition through zero-shot learning. In: WACV (2018)
26. Shrestha, R., Kafle, K., Kanan, C.: Answer them all! toward universal visual question answering models. In: CVPR (2019)
27. Wan, B., Zhou, D., Liu, Y., Li, R., He, X.: Pose-aware multi-level feature network for human object interaction detection. In: ICCV (2019)
28. Wan, H., Luo, Y., Peng, B., Zheng, W.S.: Representation learning for scene graph completion via jointly structural and visual embedding. In: IJCAI (2018)
29. Wang, T., Anwer, R.M., Khan, M.H., Khan, F.S., Pang, Y., Shao, L., Laaksonen, J.: Deep contextual attention for human-object interaction detection. In: ICCV (2019)
30. Wang, W., Wang, R., Shan, S., Chen, X.: Exploring context and visual pattern of relationship for scene graph generation. In: CVPR (2019)
31. Xu, B., Li, J., Wong, Y., Zhao, Q., Kankanhalli, M.S.: Interact as you intend: Intention-driven human-object interaction detection. IEEE Transactions on Multimedia pp. 1–1 (2019). https://doi.org/10.1109/TMM.2019.2943753
32. Xu, B., Wong, Y., Li, J., Zhao, Q., Kankanhalli, M.S.: Learning to detect human-object interactions with knowledge. In: CVPR (2019)
33. Yang, X., Tang, K., Zhang, H., Cai, J.: Auto-encoding scene graphs for image captioning. In: CVPR (2019)
34. Yao, T., Pan, Y., Li, Y., Mei, T.: Hierarchy parsing for image captioning. arXiv preprint arXiv:1909.03918 (2019)
35. Zhou, P., Chi, M.: Relation parsing neural network for human-object interaction detection. In: ICCV (2019)