PPDM: Parallel Point Detection and Matching for Real-time Human-Object Interaction Detection

Yue Liao[1,2] Si Liu[1]* Fei Wang[2] Yanjie Chen[2] Chen Qian[2] Jiashi Feng[3]
[1] School of Computer Science and Engineering, Beihang University
[2] SenseTime Research [3] National University of Singapore

Abstract

We propose a single-stage Human-Object Interaction (HOI) detection method that has outperformed all existing methods on HICO-DET dataset at 37 fps on a single Titan XP GPU. It is the first real-time HOI detection method. Conventional HOI detection methods are composed of two stages, i.e., human-object proposals generation, and proposals classification. Their effectiveness and efficiency are limited by the sequential and separate architecture. In this paper, we propose a Parallel Point Detection and Matching (PPDM) HOI detection framework. In PPDM, an HOI is defined as a point triplet < human point, interaction point, object point>. Human and object points are the center of the detection boxes, and the interaction point is the midpoint of the human and object points. PPDM contains two parallel branches, namely point detection branch and point matching branch. The point detection branch predicts three points. Simultaneously, the point matching branch predicts two displacements from the interaction point to its corresponding human and object points. The human point and the object point originated from the same interaction point are considered as matched pairs. In our novel parallel architecture, the interaction points implicitly provide context and regularization for human and object detection. The isolated detection boxes unlikely to form meaningful HOI triplets are suppressed, which increases the precision of HOI detection. Moreover, the matching between human and object detection boxes is only applied around limited numbers of filtered candidate interaction points, which saves much computational cost. Additionally, we build a new application-oriented database named as HOI-A, which serves as a good supplement to the existing datasets[1].

1. Introduction

Human-Object Interaction (HOI) detection [30, 11, 10, 9, 12, 16, 22] has received increasing attention recently.

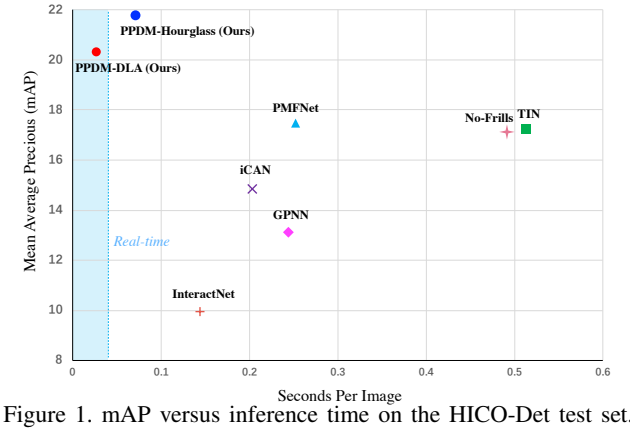

Figure 1. mAP versus inference time on the HICO-Det test set. Our PPDM-DLA outperforms the state-of-the-art methods with the inference speed of 37 fps (0.027s). It is the first real-time HOI detection method. Our PPDM-Hourglass achieves 4.27% mAP improvement over the state-of-the-arts with a faster speed.

Given an image, HOI detection aims to detect the triplet < human, interaction, object >. Different from the general visual relationship detection [19, 29, 20, 13, 32], the subject of the triplet is fixed as human while the interaction is action. HOI detection is an important step toward the high-level semantic understanding of human-centric scenes. It has a lot of applications, such as activity analysis, human-machine interaction and intelligent monitoring.

The conventional HOI detection methods [2, 22, 12, 16, 26] mostly consist of two stages. The first stage is the human-object proposal generation. A pre-trained detector [8, 23] is used to localize both the humans and objects. Then $M \times N$ human-object proposals are generated by pairwisely combining the filtered M human boxes and N object boxes. The second stage is the proposal classification which predicts the interactions for each human-object proposal. The limitations of the two-stage methods' effectiveness and efficiency are mainly because their two stages are *sequential and separated*. The proposal generation stage is completely based on object detection confidences. Each human/object proposal is independently generated. The possibility of combining two proposals to form a meaningful

*Corresponding author (liusi@buaa.edu.cn)
[1]https://github.com/YueLiao/PPDM

HOI triplet in the second stage is not taken into account. Therefore, the generated human-object proposals may have relatively low quality. Moreover, in the second stage, all human-object proposals need to be linearly scanned, while only a few of them are valid. The extra computational cost is large. Therefore, we argue that the *non-sequential and highly-coupled* framework is needed.

We propose a *parallel* HOI detection framework and reformulate HOI detection as a point detection and matching problem. As shown in Figure 2, we represent a box as a center point and corresponding sizes (width and height). Moreover, we define an interaction point as the midpoint of the human and object center points. To match each interaction point with the human point and the object point, we design two displacements from the interaction point to the corresponding human and object point. Based on the novel reformulation, we design a novel single-stage framework Parallel Point Detection and Matching (PPDM), which breaks up the complex task of HOI detection into two simpler parallel tasks. The PPDM is composed of two parallel branches. The first branch is *points detection*, which estimates the three center points (interaction, human and object points), corresponding sizes (width and height) and two local offsets (human and object points). The interaction point can be considered as providing contextual information for both human and object detection. In other words, estimating the interaction point implicitly enhances the detection of humans and objects. The second branch is *points matching*. Two displacements from the interaction point to human and object points are estimated. The human and object points originated from the same interaction points are considered as matched. In the novel parallel architecture, the point detection branch estimates the interaction points, which implicitly provide context and regularization for the human and object detection. The isolated detection boxes unlikely to form meaningful HOI triplets are suppressed while the more likely detection boxes are enhanced. It is different from the human-object proposal generation stage in two-stage methods, where all detection human/object boxes indiscriminately form the human-object proposals to feed into the second stage. Moreover, in the point matching branch, the matching is only applied around limited numbers of filtered candidate interaction points, which saves a lot of computational costs. On the contrary, in the proposal classification stage of two-stage methods, all human-object proposals need to be classified. Experimental results on the public benchmark HICO-Det [2] and our newly collected HOI-A dataset show that PPDM outperforms state-of-the-art methods in terms of accuracy and speed.

The existing datasets such as HICO-Det [22] and V-COCO [11] have greatly boosted the development of related research. These datasets are very general. However, in practical applications, several *limited, frequent* HOI categories need to be paid special attention to. To this end, we collect a new Human-Object Interaction for Applications dataset (HOI-A) with the following features: 1) specially selected 10 kinds of HOI categories with wide application values, such as smoke and ride. 2) huge intra-class variations including various illuminations and different human poses for each category. The HOI-A is more application-driven, severing as a good supplement to the existing datasets.

Figure 2. PPDM contains two parallel branches. In the point detection branch, the human/object box denoted as the center points, widths, and heights, are detected. Moreover, an interaction point, i.e., the midpoint of the human and object point, is also localized. Simultaneously, in the point matching branch, two displacements from each interaction point to the human/object are estimated. The human point and the object point originated from the same interaction point are considered as matched pairs.

Our contributions are summarized as: 1) We reformulate the HOI detection task as a point detection and matching problem and propose a novel one-stage PPDM solution. 2) PPDM is the first HOI detection method to achieve real-time and outperform state-of-the-art methods both HICO-Det and HOI-A benchmarks. 3) A large-scale and application-oriented HOI detection dataset is collected to supplement existing datasets. Both source code and the dataset are to be released to facilitate the related research.

2. Related Work

HOI Detection Methods. The existing HOI detection methods can be mostly divided into two stages: in the first stage, an object detector [23] is applied to localize the human and objects; in the second stage, pairing the detected human and object, and feeding their features into a classification network to predict the interaction between the human and object. Current works pay more attention to exploring how to improve the second stage. The most recent works aim to understand HOI by capturing context information [7, 27] or human structural message [26, 6, 5, 33]. Some works [22, 28, 33] formulated the second stage as a graph reasoning problem and use graph convolutional network to predict the HOI.

The above methods are all proposal based, thus their performance is limited by the quality of proposals. Addi-

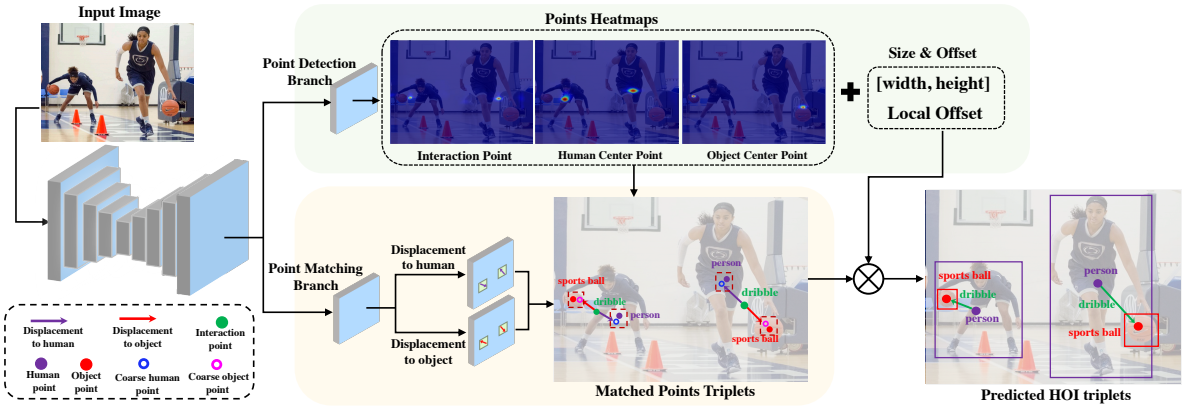

Figure 3. Overview of the proposed PPDM framework. We firstly apply a key-point heatmap prediction network, e.g. Hourglass-104 or DLA-34, to extract the appearance feature from an image. a) Point Detection Branch: Based on the extracted visual feature, we utilize three convolutional modules to predict the heatmap of the interaction points, human center points, and object center points. Additionally, to generate the final box, we regress the 2-D size and the local offset. b) Point Matching Branch: the first step of this branch is to regress the displacements from the interaction point to the human point and object point respectively. Based on the predicted points and displacements, the second step is to match each interaction point with the human point and object point to generate a set of points triplets.

tionally, the existing methods have to spend much computational cost in proposals generation and feature extraction process. Based on these drawbacks, we propose a novel one-stage and proposal-free framework to detect HOI.

HOI Detection Datasets. There are mainly two commonly used HOI detection benchmarks: VCOCO [11] and HICO-Det [2], and a human-centric relationship detection dataset: HCVRD [36]. The VCOCO is a relatively small dataset, which is a subset of MSCOCO [18] including $10,346$ images annotated with 26 actions based on COCO annotation. The HICO-Det is a large-scale and generic HOI detection dataset, including $47,776$ images, which has 117 verbs and 80 object categories (same as COCO). The HCVRD is collected from the general visual relationship detection dataset, Visual Genome [14]. It has $52,855$ images, 927 predicate categories and $1,824$ kinds of objects. Comparing the former two HOI detection datasets, which only focuses on human actions, the HCVRD is concerned about a more general human-centric relationship, e.g., spatial relationships, possessive relationships.

The previous HOI detection datasets mostly concentrate on common and general actions. From a practical view, we build up a new HOI-A dataset, which has about 38K images only annotated with limited typical kinds of actions with practical significance.

3. Parallel Point Detection and Matching

3.1. Overview

HOI detection aims to estimate the HOI triplet $<$ human, interaction, object $>$, which is composed of the subject box and class, the human action class and the object box and class. We break up the complex task of HOI detection into two simpler parallel tasks that can be assembled to form the final results. The framework of the proposed Parallel Point Detection and Matching (PPDM) method is shown in Figure 3. The first branch of PPDM is *points detection*. It estimates the center points, corresponding sizes (width and height) and local offsets of both humans and objects. The center, size and offset collaboratively represent some box candidates. Moreover, the interaction point which is defined as the midpoint of a corresponding $<$ human center point, object center point $>$ pair is also estimated. The second branch of PPDM is *points matching*. The displacements between the interaction point and the corresponding human and object points are estimated. The human point and the object point originated from the same interaction point are considered as matched pairs.

3.2. Point Detection

The point detection branch estimates the human box, object box and interaction point. A human box is represented as its center point $(x^h, y^h) \in \mathbb{R}^2$, the corresponding size (width and height) $(w^h, h^h) \in \mathbb{R}^2$ as well as the local point offset $\boldsymbol{\delta c^h} \in \mathbb{R}^2$ to recover the discretization error caused by the output stride. The object box is represented similarly. Moreover, we define the interaction point $(x^a, y^a) \in \mathbb{R}^2$ as the midpoint of the paired human point and object point. Considering the receptive field of the interaction point is large enough to contain both human and object, the human action a can be estimated based on the feature of (x^a, y^a). Actually, when there are M human in the dataset, each human box is represented as $(x_i^h, y_i^h), i \in [1, M]$. For the convenience of description, we omit the subscript i when no confusion is caused. Similar omissions are also applicable for (x^o, y^o) and (x^a, y^a).

In Figure 3, the input image $\boldsymbol{I} \in \mathbb{R}^{H \times W}$ is fed into the feature extractor to produce the feature $\boldsymbol{V} \in \mathbb{R}^{\frac{H}{d} \times \frac{W}{d}}$,

where W and H are the width and height of the input image and d is the output stride. The point heatmaps are of low-resolution, thus we also calculate the low-resolution center points. Given a ground-truth human point (x^h, y^h), the corresponding low-resolution point is $(\tilde{x}^h, \tilde{y}^h) = (\lfloor \frac{x^h}{d} \rfloor, \lfloor \frac{y^h}{d} \rfloor)$. The low-resolution ground-truth object point $(\tilde{x}^o, \tilde{y}^o)$ can be computed in the same way. Based on the low-resolution human and object points, the ground-truth interaction point can be defined as $(\tilde{x}^a, \tilde{y}^a) = (\lfloor \frac{\tilde{x}^h + \tilde{x}^o}{2} \rfloor, \lfloor \frac{\tilde{y}^h + \tilde{y}^o}{2} \rfloor)$.

Point location loss. Directly detecting a point is difficult, thus we follow the key-point estimation methods [25] to splat a point into a heatmap with a Gaussian kernel. Thereby the point detection is transformed into a heatmap estimation task. The three ground-truth low-resolution points (x^h, y^h), (x^o, y^o) and (x^a, y^a) are splatted into three Gaussian heatmaps, including human point heatmap $\tilde{\boldsymbol{C}}^{\boldsymbol{h}} \in [0,1]^{\frac{H}{d} \times \frac{W}{d}}$, object point heatmap $\tilde{\boldsymbol{C}}^{o} \in [0,1]^{T \times \frac{H}{d} \times \frac{W}{d}}$ and interaction point heatmap $\tilde{\boldsymbol{C}}^{a} \in [0,1]^{K \times \frac{H}{d} \times \frac{W}{d}}$, where T is the number of object categories and K is the the number of interaction classes. Note that in $\tilde{\boldsymbol{C}}^{\boldsymbol{o}}$ and $\tilde{\boldsymbol{C}}^{\boldsymbol{a}}$, only the channel corresponding to the specific object class and human action are non-zero. The three heatmaps are produced by adding three respective convolutional blocks upon the feature map $\boldsymbol{V}$, each of which is composed of a 3×3 convolutional layer with ReLU, followed by a 1×1 convolutional layer and a Sigmoid.

For the three heatmaps, we all apply an element-wise focal loss [17]. For example, given an estimated interaction point heatmap $\hat{\boldsymbol{C}}^{\boldsymbol{a}}$ and the corresponding ground-truth heatmaps $\tilde{\boldsymbol{C}}^{\boldsymbol{a}}$, the loss function is:

$$L_a = -\frac{1}{N} \sum_{kxy} \begin{cases} (1 - \hat{C}^a_{kxy})^{\alpha} \log(\hat{C}^a_{kxy}) & \text{if } \tilde{C}^a_{kxy} = 1 \\ (1 - \tilde{C}^a_{kxy})^{\beta} (\hat{C}^a_{kxy})^{\alpha} & \text{otherwise} \\ \log(1 - \hat{C}^a_{kxy}), & \end{cases} \tag{1}$$

where N is equal to the number of interaction points (HOI triplet) in the image, and $\hat{C}^a_{kxy}$ is the score at location (x, y) for class k in the predicted heatmaps $\hat{\boldsymbol{C}}^{\boldsymbol{a}}$. We set α as 2 and β as 4 following the default setting in [15, 35, 4]. The losses L_p and L_o for the human points and the object points can be computed similarly.

Size and offset loss. Besides the center points, the box size and the local offset for the center points are needed to form the human/object box. Four convolutional blocks are added to the feature map $\boldsymbol{V}$ to estimate the 2-D size and the local offset of human and object boxes respectively. Each block contains a 3×3 convolutional layer with ReLU and a 1×1 convolutional layer.

During training, we only compute the $L1$ loss at each location of the ground truth human point $(\tilde{x}^h, \tilde{y}^h)$ and object point $(\tilde{x}^o, \tilde{y}^o)$ and ignore all other locations. We take the loss function for the local offset as an example, while the size regression loss L_{wh} is defined similarly. The ground truth local offset for the human point localized at $(\tilde{x}^h, \tilde{y}^h)$ is defined as $(\tilde{\delta}^x_{(\tilde{x}^h, \tilde{y}^h)}, \tilde{\delta}^y_{(\tilde{x}^h, \tilde{y}^h)}) = (\frac{x^h}{d} - \tilde{x}^h, \frac{y^h}{d} - \tilde{y}^h)$. Thus the loss function L_{off} is the summation of the human box loss L^h_{off} and object box loss L^o_{off}.

$$L_{off} = \frac{1}{M + D}(L^h_{off} + L^o_{off}) \tag{2}$$

$$\begin{aligned} L^h_{off} = \sum_{(\tilde{x}^h, \tilde{y}^h) \in \tilde{S}^h} & (|\tilde{\delta}^x_{(\tilde{x}^h, \tilde{y}^h)} - \hat{\delta}^x_{(\tilde{x}^h, \tilde{y}^h)}| \\ & + |\tilde{\delta}^y_{(\tilde{x}^h, \tilde{y}^h)} - \hat{\delta}^y_{(\tilde{x}^h, \tilde{y}^h)}|, \end{aligned} \tag{3}$$

where $\tilde{S}^h$ and $\tilde{S}^h$ denote the ground-truth human and object points sets in the training set. $M = |\tilde{S}^h|$ and $D = |\tilde{S}^o|$ are the number of human points and object points. Note that M is not necessarily equal to D. For example a human may correspond to multiple actions and objects. L^o_{off} is defined similarly with Equation 3.

3.3. Point Matching

The points matching branch pairs the human box with its corresponding object box by using the interaction point as the bridge. More specifically, the interaction point is treated as the anchor. Two displacements $\boldsymbol{d}^{\boldsymbol{ah}} = (d^{ah}_x, d^{ah}_y)$ and $\boldsymbol{d}^{\boldsymbol{ao}} = (d^{ao}_x, d^{ao}_y)$, i.e., the displacements between interaction point vs. human/box point are estimated. The coarse human point and object point are (x^a, y^a) plus $\boldsymbol{d}^{\boldsymbol{ah}}$ and $\boldsymbol{d}^{\boldsymbol{ao}}$ respectively.

Our proposed displacement branch is composed of two convolutional modules. Each module consists of a 3×3 convolutional layer with ReLU and a 1×1 convolutional layer. The size of both subject and object displacement maps are $2 \times \frac{H}{d} \times \frac{W}{d}$.

Displacement loss. To train the displacement branch, we apply $L1$ loss for each interaction point. The ground-truth displacement from the interaction point located at $(\tilde{x}^a, \tilde{y}^a)$ to the corresponding human point can be computed by $(\tilde{d}^{hx}_{(\tilde{x}^a, \tilde{y}^a)}, \tilde{d}^{hy}_{(\tilde{x}^a, \tilde{y}^a)}) = (\tilde{x}^a - \tilde{x}^h, \tilde{y}^a - \tilde{y}^h)$. The predicted displacement at location of $(\tilde{x}^a, \tilde{y}^a)$ is $(\hat{d}^{hx}_{(\tilde{x}^a, \tilde{y}^a)}, \hat{d}^{hy}_{(\tilde{x}^a, \tilde{y}^a)})$. The displacement loss is defined as:

$$\begin{aligned} L_{ah} = & \frac{1}{N} \sum_{(\tilde{x}^a, \tilde{y}^a) \in \tilde{S}^a} |\hat{d}^{hx}_{(\tilde{x}^a, \tilde{y}^a)} - \tilde{d}^{hx}_{(\tilde{x}^a, \tilde{y}^a)}| \\ & + |\hat{d}^{hx}_{(\tilde{x}^a, \tilde{y}^a)} - \tilde{d}^{hy}_{(\tilde{x}^a, \tilde{y}^a)}|, \end{aligned} \tag{4}$$

where $\tilde{S}^a$ denotes the ground-truth interaction point sets in the training set. $N = |\tilde{S}^a|$ is the number of interaction points. The loss function for displacement from the interaction point to the object point L_{ao} has the same form.

Triplet matching. Two aspects are considered to judge whether a human/object point can be matched with the interaction point. The human/object needs to: 1) be close

to the coarse human/object point generated by interaction point plus the displacement and 2) have high confidence scores. On basis of these, for the detected interaction point $(\hat{x}^a, \hat{y}^a)$, we rank the points in the detected human point set $\hat{S}^h$ by Equation 5 and select the optimal one.

$$(\hat{x}^h_{opt}, \hat{y}^h_{opt}) = \underset{(\hat{x}^h,\hat{y}^h)\in\hat{S}^h}{\arg\min} \frac{1}{C^h_{(\hat{x}^h,\hat{y}^h)}} \cdot (|(\hat{x}^a, \hat{y}^a) - (\hat{d}^{hx}_{(\hat{x}^a,\hat{y}^a)}, \hat{d}^{hy}_{(\hat{x}^a,\hat{y}^a)}) - (\hat{x}^h, \hat{y}^h)|) \quad (5)$$

where $C^h_{(\hat{x}^h,\hat{y}^h)}$ denotes the confidence score for human point $(\hat{x}^h, \hat{y}^h)$. The optimal object box $(\hat{x}^o_{opt}, \hat{y}^o_{opt})$ can be calculated similarly.

3.4. Loss and Inference

The final loss can be obtained by weighted summing the above losses:

$$L = L_a + L_h + L_o + \lambda(L_{ah} + L_{ao} + L_{wh}) + L_{off} \quad (6)$$

where we set the λ as 0.1 following [15, 35]. L_a, L_h and L_o are point location loss, L_{ah} and L_{oh} are displacement loss while L_{wh} and L_{off} are size and offset lose .

During the inference, we firstly do a 3×3 max-pooling operation with stride 1 on the predicted human, object and interaction points heatmap, which plays a similar role as NMS. Secondly, we select top K human points $\hat{S}^h$, object center points $\hat{S}^o$ and interaction points $\hat{S}^a$ through the corresponding confidence scores $\hat{C}^h$, $\hat{C}^o$ and $\hat{C}^a$ across all categories. Then, we find the subject point and object point for each selected interaction point by Equation 5. For each matched human point $(\hat{x}^h_{opt}, \hat{y}^h_{opt})$, we get the final box as:

$$\begin{aligned}(\hat{x}^h_{ref} - \frac{\hat{w}_{(\hat{x}^h_{opt},\hat{y}^h_{opt})}}{2}, \hat{y}^h_{ref} - \frac{\hat{h}_{(\hat{x}^h_{opt},\hat{y}^h_{opt})}}{2}, \\ \hat{x}^h_{ref} + \frac{\hat{w}_{(\hat{x}^h_{opt},\hat{y}^h_{opt})}}{2}, \hat{y}^h_{ref} + \frac{\hat{h}_{(\hat{x}^h_{opt},\hat{y}^h_{opt})}}{2}.\end{aligned} \quad (7)$$

where $\hat{x}^h_{ref} = \hat{x}^h_{opt} + \hat{\delta}^x_{(\hat{x}^h_{opt},\hat{y}^h_{opt})}$ and $\hat{y}^h_{ref} = \hat{y}^h_{opt} + \hat{\delta}^y_{(\hat{x}^h_{opt},\hat{y}^h_{opt})}$ are the refined location of the human center point. $(\frac{\hat{w}_{(\hat{x}^h_{opt},\hat{y}^h_{opt})}}{2}, \frac{\hat{h}_{(\hat{x}^h_{opt},\hat{y}^h_{opt})}}{2})$ is the size of box in the corresponding position. The final HOI detection results are a set of triplets, and the confidence score for the triplet is $\hat{C}^p_{\hat{x}^h_{ref}\hat{y}^h_{ref}} \hat{C}^o_{\hat{x}^o_{ref}\hat{y}^o_{ref}} \hat{C}^a_{\hat{x}^a_{ref}\hat{y}^a_{ref}}$.

4. HOI-A Dataset

The existing datasets such as HICO-Det [22] and V-COCO [11] have greatly boosted the development of related research. However, in practical application, there are limited frequent HOI categories that need to be paid special attention to, which are not emphasized in previous datasets.

Figure 4. Example images of our HOI-A dataset. We take <human, smoke, cigarette> as an example. The (a)-(d) show huge intra-class variations of <human, smoke, cigarette> in the wild. The (e)-(f) show two kinds of negative samples.

We then introduce a new dataset called Human-Object Interaction for Application (HOI-A).

As shown in Table 1, we select the categories of verb driven by practical application. Each kind of verb in HOI-A dataset has its corresponding application scenario, for example 'talk on' can be applied in dangerous action detection, e.g., if the human is talking on phone in-car, it can be considered as a dangerous driving action.

Verbs	Objects	# Instance
smoking	cigarette	8624
talk on	mobile phone	18763
play (mobile phone)	mobile phone	6728
eat	food	831
drink	drink	6898
ride	bike, motorbike, horse	7111
hold	cigarette, mobile phone, food drink, document, computer	44568
kick	sports ball	365
read	document	869
play (computer)	computer	1402

Table 1. The list and occurrence numbers of the verbs of the corresponding objects in HOI-A dataset.

4.1. HOI-A Construction

We describe the image collection and annotation process for constructing the HOI-A dataset. The first step is collecting candidate images, which can be divided into two parts, namely positive and negative images collections.

Positive Images Collection. We collect positive images in two ways, i.e., camera shooting and crawling. Camera shooting is an important way to enlarge the intra-class variations of the data. We employ 50 performers and require them to perform all predefined actions in different scenes and illumination, with various poses, and take photos of them respectively with an RGB camera and an IR camera. For data crawling from the Internet, we generate a series of keywords based on the HOI triplet < person, action name, object name>, action pair <action name, object name> and action name, and retrieve images from the Internet.

Negative Images Collection. Negative Images Collection. There are two kinds of negative samples of the predefined

$<$ human, interaction, object $>$. 1) The concerned object appears in the image, but the concerned action does not happen. For example, in Figure 4(f), although the cigarette appears in the image, it is not smoked by a human. Therefore, the image is still a negative sample. 2) Other action similar to the concerned action happens but the concerned object is missing. For example, in Figure 4(e), the man is smoking at a glance. But a closer look shows there is no cigarette in the image. We collect this kind of negative sample in the 'attack' manner. We firstly train a multi-label action classifier based on the annotated positive images. The classifier takes an image as input and outputs the probability of action classification. Then, we let actors perform arbitrarily to attack the classifier without any interacted objects. If the attack is successful, we record this image as a hard negative sample.

Annotation. The process of annotation contains two steps: box annotation and interaction annotation. First, all objects in the pre-defined categories are annotated with a box and the corresponding category. Second, we visualize the boxes in the images with their id and annotate whether a person has the defined interactions with a object. The annotator should record the triplet <person ID, action ID, object ID>. For more accurate annotation, each image is annotated by 3 annotators. The annotation of an image is regarded qualified if at least 2 annotators share the same annotation.

4.2. Dataset Properties

Scale. Our HOI-A dataset consists of 38,668 annotated images, 11 kinds of objects and 10 action categories. In detail, it contains $43,820$ human instances, $60,438$ object instances and $96,160$ interaction instances. There are on average 2.2 interactions performed per person. Table 1 lists the number of instances for each verb which occurs at least 360 times. 60% verbs appear more than $6,500$ times. To our knowledge, this is already the largest HOI dataset, in terms of the number of images per interaction category.

Intra-Class Variation. To enlarge the intra-class variation of the data, each type of verbs in our HOI-A dataset will be captured with three general scenes including indoor, outdoor and in-car, three lighting conditions including dark, natural and intense, various human poses and different angles. Additionally, we shoot the images with two kinds of cameras: RGB and IR.

5. Experiments

5.1. Experimental Setting

Datasets. To verify the effectiveness of our PPDM, we conduct experiments not only on our HOI-A dataset but also on the general HOI detection dataset HICO-Det [2]. HICO-Det is a large-scale dataset for common HOI detection. It has $47,776$ images ($38,118$ for training and $9,658$ for test), annotated with 117 verbs including 'no-interaction' and 80 object categories. The 117 verbs and 80 objects form 600 kinds of HOI triplets, where 138 types of HOIs which appear <10 times are considered as the rare set, and the rest 462 kinds of HOIs form the non-rare set.

Metric. Following the standard-setting in HOI detection task, we use mean average precious (mAP) as the metric. If a predicted triplet is considered as a true positive sample, it needs to match a certain ground-truth triplet. Specifically, they have the same HOI class and their human and object boxes have overlap with IOUs large than 0.5. There is a slight difference when computing AP on the two datasets. We compute AP per HOI class in HICO-Det and compute AP per verb class in HOI-A dataset.

Implementation Details. We use two common heatmap prediction networks as our feature extractor, Hourglass-104 [21, 15] and DLA-34 [31, 35]. Hourglass-104 is a general heatmap prediction network commonly used in keypoint detection and object detection. In PPDM, we use the modified version Hourglass-104 proposed in [15]. The DLA-34 is a lightweight backbone network, and we apply a refined version proposed in [35]. The receptive field of the network need large enough to cover the subject and the object. Hourglass-104 has a sufficiently large receptive field, while that of DLA-34 cannot cover the region including the human and the object, due to its relatively shallow architecture. Thus for the DLA-based model, we concatenate the last three level features and apply a graph-based global reasoning module [3] to enlarge the receptive field for the interaction point and displacement prediction. In the global reasoning module, we set the channels of the node and the reduced feature as 48 and 96 respectively. For Hourglass-104, we only use the last-level feature for all the following modules. We initialize the feature extractor with the weights pre-trained on COCO [18]. Our experiments are all conducted on the Titan Xp GPU and CUDA 9.0.

During training and inference, the input resolution is 512×512 and the output is 128×128. PPDM is trained with Adam on 8 GPUs. We set the hyper-parameter following [35], which is robust to our framework. We train the model based on DLA-34 with a 128 sized mini-batch for 110 epochs, with a learning rate of 5e-4 decreased to 5e-5 at the 90th epoch. For the hourglass-104 based model, we train it with a batch size of 32 for 110 epochs, with a learning rate of 3.2e-4 decreased by 10 times at the 90th epoch. We follow [15, 35] applying data augmentation, i.e., random scale and random shift to train the model and there is no augmentation during inference. We set the number of selected predictions K as 100.

5.2. Comparison to State-of-the-art

We compare PPDM with state-of-the-art methods on two datasets. The quantitative results can be seen in Table 2 and Table 3, and the qualitative results are presented in Figure 5. The compared methods mainly use a pre-trained Faster R-CNN [23] to generate a set of human-object pro-

Method	Feature	Default			Know Object			Inference Time (ms) ↓	FPS ↑
		Full	Rare	Non-Rare	Full	Rare	Non-Rare		
Shen *et. al* [24]	A + P	6.46	4.24	7.12	-	-	-	-	-
HO-RCNN [2]	A + S	7.81	5.37	8.54	10.41	8.94	10.85	-	-
InteractNet [9]	A	9.94	7.16	10.77	-	-	-	145	6.90
GPNN [22]	A	13.11	9.34	14.23	-	-	-	197 + 48 = 245	4.08
Xu *et. al* [28]	A + L	14.70	13.26	15.13	-	-	-	-	-
iCAN [7]	A + S	14.84	10.45	16.15	16.26	11.33	17.73	92 + 112 = 204	4.90
PMFNet-Base [26]	A + S	14.92	11.42	15.96	18.83	15.30	19.89	-	-
Wang *et. al* [27]	A	16.24	11.16	17.75	17.73	12.78	19.21	-	-
No-Frills [12]	A + S + P	17.18	12.17	18.68	-	-	-	197 + 230 + 67 = 494	2.02
TIN [16]	A + S + P	17.22	13.51	18.32	19.38	15.38	20.57	92 + 98 + 323 = 513	1.95
RPNN [33]	A + P	17.35	12.78	18.71	-	-	-	-	-
PMFNet [26]	A + S + P	17.46	**15.65**	18.00	20.34	**17.47**	21.20	92 + 98 + 63 = 253	3.95
PPDM-DLA	A	20.29	13.06	22.45	23.09	16.14	25.17	**27**	**37.03**
PPDM-Hourglass	A	**21.73**	13.78	**24.10**	**24.58**	16.65	**26.84**	71	14.08

Table 2. Performance comparison on the HICO-DET test set. The 'A', 'P', 'S', 'L' represent the appearance feature, human pose information, the spatial feature, and the language feature, respectively.

posals, which are then fed into a pairwise classification network. As shown in Table 2, to more accurately classify the HOI, many methods use additional human pose feature or language feature.

5.2.1 Quantitative Analysis

HICO-Det. See table 2. Our PPDM-DLA and PPDM-Hourglss both outperform all previous state-of-the-art methods. Specifically, our PPDM-Hourglass achieves a significant performance gain (24.5%) comparing to the previous best method PMFNet [26]. We can see the previous methods with mAP greater than 17% all use the human pose as an additional feature, while our PPDM only uses the appearance feature. Performance of PPDM is slightly lower than PMFNet on the rare subset. However, the baseline model in PMFNet without using human pose information only achieves 11.42% mAP on the rare-set. The performance gain on the rare-set may mainly come from the additional human pose feature. The human structural information plays an important role in understanding human actions, thus we regard how to utilize human context in our framework as a significant future work.

Method	mAP (%)	Time (ms)
Faster Interaction Net [1]	56.93	-
GMVM [1]	60.26	-
C-HOI [34]	66.04	-
iCAN [7]	44.23	194
TIN [16]	48.64	501
PPDM-DLA	67.45	**27**
PPDM-Hourglass	**71.23**	71

Table 3. Performance comparison on HOI-A test set.

HOI-A. The compared methods in HOI-A dataset are composed of two parts. Firstly, we select the top-3 methods from the leaderboard of ICCV 2019 PIC challenge HOI detection track [1], which was held by us based on HOI-A dataset. Comparing to the top-1 method, C-HOI [34], which uses a very strong detector, our methods still outperform it. Secondly, we choose two open-source state-of-the-art methods, iCAN [7] and TIN [16], as the baselines on our HOI-A dataset. We first pre-train Faster R-CNN with FPN and ResNet-50 on HOI-A dataset, and then follow their original settings to train the HOI classifier. The results show our PPDM outperforms the two methods by a significant margin. Additionally, for our selected interaction types with practical significance, our PPDM can achieve high performance, which is practically applicable.

5.2.2 Qualitative Analysis

We visualize the HOI prediction with the top-3 confidence score on HICO-Det dataset based on PPDM-DLA, and compare our results with the typical two-stage method iCAN [7]. As shown in Figure 5, we select some representative failure cases of the two-stage methods. We can see iCAN tends to focus on the human/object with a high detection score but without interaction. In Figure 5(b) and Figure 5(c), due to the huge imbalance between positive/negative samples, iCAN easily produces high confidence for the 'no-interaction' type. In Figure 5(d), the person sitting on the airplane is so small that it cannot be detected. However, our PPDM can accurately predict the HOI triplets with high confidence in these cases. Because PPDM is not dependent on the proposals. Moreover, PPDM concentrates on the HOI triplets understanding.

	Method	Full	Rare	Non-Rare	Time
1	Basic Model	19.94	13.01	22.01	24
2	+ Feature Fusion	20.00	12.56	22.22	26
3	+ Global Reasoning	19.85	12.99	21.90	26
4	Union Center	18.65	12.11	20.61	27
5	PPDM-DLA	20.29	13.06	22.45	27

Table 4. Component analysis on HICO-Det Test Set.

5.2.3 Efficiency Analysis

We compare the inference speed on a single Titan Xp GPU with the methods which have released code or reported the

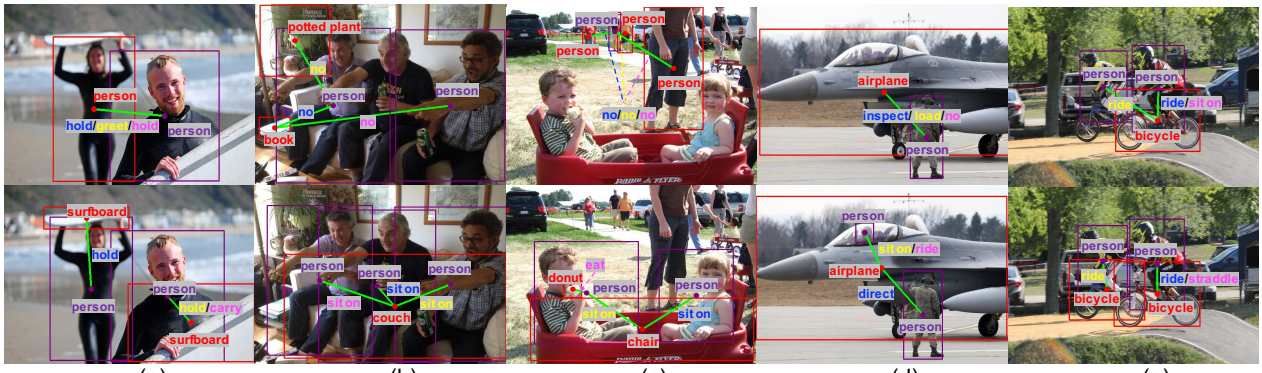

Figure 5. Visualization results compared with iCAN on HICO-Det. The first row is the prediction of iCAN and the second row by PPDM. Purple denotes the subject and red is the object. If a subject has interaction with an object, they will be linked by a green line. We show results with top-3 confidence per image: 1-blue, 2-yellow, 3-pink. The 'no' denotes 'no-interaction'.

speed. As shown in Table 2, PPDM with DLA and Hourglass are both faster than other methods by a large margin. PPDM-DLA is the only real-time method, which only takes 27ms for inference. Concretely, the inference time of two-stage HOI detection methods can be divided into proposal generation time and HOI classifier time. Besides, the pose based methods take extra time to estimate human keypoints. It can be seen that the speed of PPDM-DLA is faster than any stage of the compared methods.

5.3. Component Analysis

We analyze the proposed components in PPDM from quantitative and qualitative views.

Figure 6. Visualization of interaction points heatmaps and displacements. Red and purple line represent displacements from interaction point (green) to human and object.

Feature Extractor. We analyze effectiveness of the additional modules in DLA backbone, i.e., feature fusion and global reasoning. The first row in Table 4 represents the basic framework with DLA, where we predict the interaction only based on the last-level feature. It shows that the basic model can still outperform all existing methods. It proves the effectiveness of our designed framework. The second row and third row show the results of the basic model with the feature fusion and a global reasoning module respectively, it can be seen in Table 4 that the performance only change little. If we add a these two settings to the basic framework as the same time, that the performance improves by 0.35% mAP. We conclude that a larger receptive field and global context are helpful to interaction prediction.

Point Detection. To verify whether the midpoint of two center points is the best choice to predict the interaction, we perform an experiment based on the interaction point at the center of the union of human and object boxes, which is another suitable location to predict the interaction. See the 4th row in Table 4. The mAPs drop 1.64 point compared with PPDM-DLA. It is common that two objects interact with the same person and may locate in the human box, in which case the center points of their union boxes overlap. Additionally, we analyze our interaction point qualitatively. As shown in Figure 6, the predicted interaction almost accurately locates at the midpoint of the human/object points, though the human is apart from the object or in the object.

Point Matching. To further understand the displacement, we visualize the displacements in Figure 6. We can see the interaction point plus the corresponding displacement is very close to the center point of the human/object box, even though the human/object is hard to be detected.

6. Conclusion

In this paper, we propose a novel one-stage framework and a new dataset for HOI detection. Our proposed method can outperform the existing methods by a margin also with a significantly faster speed. It breaks the limits of the traditional two-stage methods and directly predicts the HOI by a parallel framework. Our proposed HOI-A dataset is more inclined to practical application for HOI detection.

Acknowledgement This work was partially supported by Sensetime Ltd. Group, Zhejiang Lab (No. 2019KD0AB04), Beijing Natural Science Foundation (L182013, 4202034) and Fundamental Research Funds for the Central Universities.

References

[1] Pic leaderboard. http://www.picdataset.com/challenge/leaderboard/hoi2019.

[2] Yu-Wei Chao, Yunfan Liu, Xieyang Liu, Huayi Zeng, and Jia Deng. Learning to detect human-object interactions. In *WACV*, 2018.

[3] Yunpeng Chen, Marcus Rohrbach, Zhicheng Yan, Yan Shuicheng, Jiashi Feng, and Yannis Kalantidis. Graph-based global reasoning networks. In *CVPR*, 2019.

[4] Zhiwei Dong, Guoxuan Li, Yue Liao, Fei Wang, Pengju Ren, and Chen Qian. Centripetalnet: Pursuing high-quality keypoint pairs for object detection. In *CVPR*, 2020.

[5] Hao-Shu Fang, Jinkun Cao, Yu-Wing Tai, and Cewu Lu. Pairwise body-part attention for recognizing human-object interactions. In *ECCV*, 2018.

[6] Wei Feng, Wentao Liu, Tong Li, Jing Peng, Chen Qian, and Xiaolin Hu. Turbo learning framework for human-object interactions recognition and human pose estimation. 2019.

[7] Chen Gao, Yuliang Zou, and Jia-Bin Huang. ican: Instance-centric attention network for human-object interaction detection. In *BMVC*, 2018.

[8] Ross Girshick. Fast r-cnn. In *CVPR*, 2015.

[9] Georgia Gkioxari, Ross Girshick, Piotr Dollár, and Kaiming He. Detecting and recognizing human-object interactions. In *CVPR*, 2018.

[10] Abhinav Gupta, Aniruddha Kembhavi, and Larry S Davis. Observing human-object interactions: Using spatial and functional compatibility for recognition. *TPAMI*, 2009.

[11] Saurabh Gupta and Jitendra Malik. Visual semantic role labeling. *arXiv preprint arXiv:1505.04474*, 2015.

[12] Tanmay Gupta, Alexander Schwing, and Derek Hoiem. No-frills human-object interaction detection: Factorization, appearance and layout encodings, and training techniques. In *ICCV*, 2019.

[13] Roei Herzig, Moshiko Raboh, Gal Chechik, Jonathan Berant, and Amir Globerson. Mapping images to scene graphs with permutation-invariant structured prediction. In *NIPS*, 2018.

[14] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73, 2017.

[15] Hei Law and Jia Deng. Cornernet: Detecting objects as paired keypoints. In *ECCV*, 2018.

[16] Yong-Lu Li, Siyuan Zhou, Xijie Huang, Liang Xu, Ze Ma, Hao-Shu Fang, Yan-Feng Wang, and Cewu Lu. Transferable interactiveness prior for human-object interaction detection. In *CVPR*, 2019.

[17] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *CVPR*, 2017.

[18] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014.

[19] Cewu Lu, Ranjay Krishna, Michael Bernstein, and Li Fei-Fei. Visual relationship detection with language priors. In *ECCV*, 2016.

[20] Alejandro Newell and Jia Deng. Pixels to graphs by associative embedding. In *NIPS*, 2017.

[21] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *ECCV*, 2016.

[22] Siyuan Qi, Wenguan Wang, Baoxiong Jia, Jianbing Shen, and Song-Chun Zhu. Learning human-object interactions by graph parsing neural networks. In *ECCV*, 2018.

[23] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*, 2015.

[24] Liyue Shen, Serena Yeung, Judy Hoffman, Greg Mori, and Li Fei-Fei. Scaling human-object interaction recognition through zero-shot learning. In *WACV*, 2018.

[25] Jonathan J Tompson, Arjun Jain, Yann LeCun, and Christoph Bregler. Joint training of a convolutional network and a graphical model for human pose estimation.

[26] Bo Wan, Desen Zhou, Yongfei Liu, Rongjie Li, and Xuming He. Pose-aware multi-level feature network for human object interaction detection. In *ICCV*, 2019.

[27] Tiancai Wang, Rao Muhammad Anwer, Muhammad Haris Khan, Fahad Shahbaz Khan, Yanwei Pang, Ling Shao, and Jorma Laaksonen. Deep contextual attention for human-object interaction detection. In *ICCV*, 2019.

[28] Bingjie Xu, Yongkang Wong, Junnan Li, Qi Zhao, and Mohan S. Kankanhalli. Learning to detect human-object interactions with knowledge. In *CVPR*, 2019.

[29] Danfei Xu, Yuke Zhu, Christopher B Choy, and Li Fei-Fei. Scene graph generation by iterative message passing. In *CVPR*, 2017.

[30] Bangpeng Yao and Li Fei-Fei. Recognizing human-object interactions in still images by modeling the mutual context of objects and human poses. *TPAMI*, 2012.

[31] Fisher Yu, Dequan Wang, Evan Shelhamer, and Trevor Darrell. Deep layer aggregation. In *CVPR*, 2018.

[32] Ji Zhang, Kevin J Shih, Ahmed Elgammal, Andrew Tao, and Bryan Catanzaro. Graphical contrastive losses for scene graph parsing. In *CVPR*, 2019.

[33] Penghao Zhou and Mingmin Chi. Relation parsing neural network for human-object interaction detection. In *ICCV*, 2019.

[34] Tianfei Zhou, Wenguan Wang, Siyuan Qi, Haibin Ling, and Jianbing Shen. Cascaded human-object interaction recognition. In *CVPR*, 2020.

[35] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. Objects as points. *arXiv preprint arXiv:1904.07850*, 2019.

[36] Bohan Zhuang, Qi Wu, Chunhua Shen, Ian Reid, and Anton van den Hengel. Care about you: towards large-scale human-centric visual relationship detection. *arXiv preprint arXiv:1705.09892*, 2017.