PPDM++: Parallel Point Detection and Matching for Fast and Accurate HOI Detection

Yue Liao, Si Liu, Yulu Gao, Aixi Zhang, Zhimin Li, Fei Wang, Bo Li

Abstract—Human-Object Interaction (HOI) detection aims to understand human activities by detecting interaction triplets. Previous HOI detection methods adopt a two-stage instance-driven paradigm. Unfortunately, many non-interactive human-object pairs generated by the first stage are the main obstacle impeding HOI detectors from high efficiency and promising performance. To remedy this, we propose a novel top-down interaction-driven paradigm, detecting interactions first and bridging interactive human-object pairs through interactions. We formulate HOI as a point triplet <human point, interaction point, object point> and design a Parallel Point Detection and Matching (PPDM) framework. We further take advantage of two-stage methods and propose a novel framework, PPDM++, that detects the interactive human-object pairs by PPDM, then extracts region features for each pair to predict actions. The core of PPDM/PPDM++ is to convert the instance-driven bottom-up paradigm to an interaction-driven top-down paradigm, thus avoiding additional computation costs from traversing a tremendous number of non-interactive pairs. Benefiting from the advanced paradigm, PPDM/PPDM++ has achieved significant performance gains with high efficiency. PPDM-DLA-34 has achieved 19.94 mAP with 42 FPS as the first real-time HOI detector, and PPDM++-SwinB achieves 30.1 mAP with 17 FPS on HICO-DET dataset. We also built an application-oriented database named HOI-A, a supplement to the existing datasets.

Index Terms—Human-Object Interaction Detection, Visual Relationship Detection, One-stage Detector, Dataset.

1 Introduction

Intelligent human activities analysis for human-centric visual scenarios is a fundamental task in the computer vision area. Human-object Interaction (HOI) detection concentrates on instance-level human interactive activities analysis in a static image. In specific, the goal of the HOI detection task is to detect interactive human and object pairs and classify their interactive actions in an image, thus generating a series of HOI triplets <Human Box, Object Box, Action>. HOI detection is considered a structural understanding of human-centric scenarios, which is an important step toward the high-level semantic understanding tasks, *e.g.*, image caption, and visual question answer. Moreover, HOI detection takes is able to support a broad range of practical applications, such as dangerous activity alerts, smart retail, and human-machine interaction.

With the development of deep learning, HOI detection [48], [15], [14], [13], [16], [27], [37] has attracted increasing attention recently. As shown in Figure 2(a), previous HOI detection methods [4], [37], [16], [27], [44], [25] are mostly built upon object detection methods with an instance-driven bottom-up mechanism and a two-stage serial architecture. Such methods first employ an instance detector [12], [38] to produce a series of human and object instances as proposals. Then, an interaction classification network is adopted to match interactive human and object instance pairs and recognize their interactive actions in a bottom-up manner. The matching and classification of conventional methods are mostly straightforward but inefficient processes, where the output instance proposals are firstly filtered to produce M human and N object proposals and matched pairwisely to form $M \times N$ human-object proposals, and then a classification module is adopted to classify the interactions of each human-object proposal. The *serial and isolated*

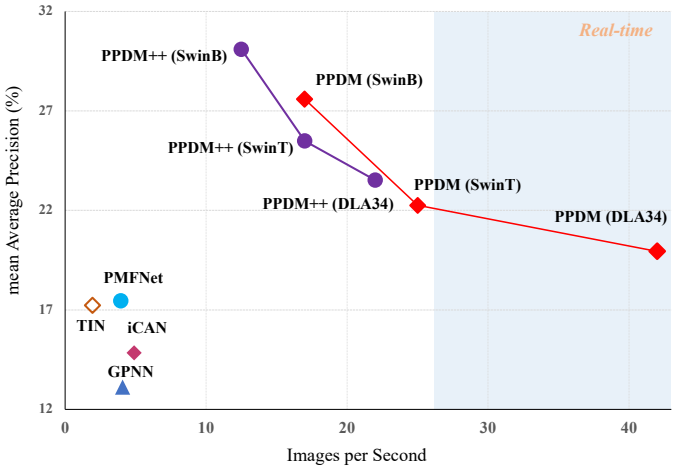

Fig. 1. mAP versus inference speed (the processed image number per second) on the HICO-DET test set. Our PPDM-DLA outperforms the traditional two-stage methods with an inference speed of 41.67 fps (0.024s). It is the first HOI detection method to achieve real-time speed. Our PPDM++ has achieved about 3 mAP improvement over PPDM in the same backbone setting. Our PPDM++ with the 'Swin-transformer base' backbone can achieve 30.1 mAP with an inference speed of 17FPS.

- *Yue Liao, Si Liu, Yulu Gao, Aixi Zhang and Bo Li are with the Institute of Artificial Intelligence, Beihang University. Email: {liaoyue.ai, zhangaixi2008}@gmail.com, {liusi,gyl97,boli}@buaa.edu.cn.*
- *Zhimin Li is with the School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, Email: zhiminli.cn@outlook.cn.*
- *Fei Wang is with the School of Information Science and Technology, University of Science and Technology of China, E-mail:wangfei91@mail.ustc.edu.cn.*

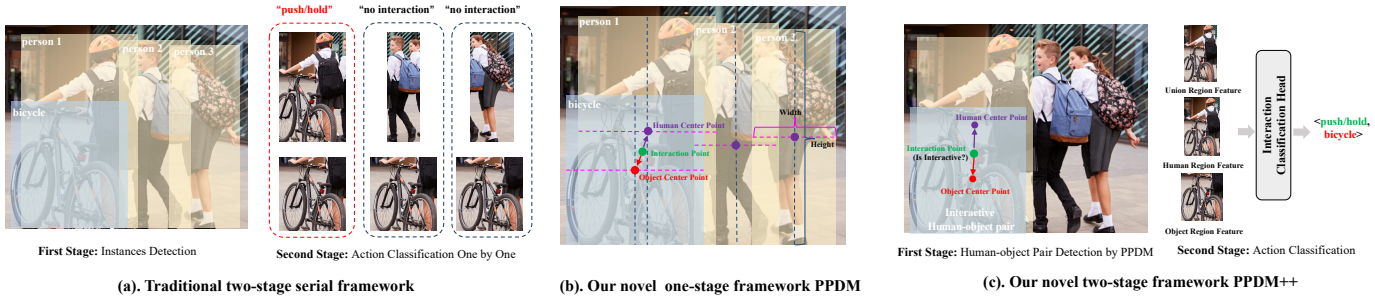

Fig. 2. The comparison of the traditional two-stage serial framework, our novel one-stage PPDM framework, and our novel two-stage framework based on PPDM, namely PPDM++. (a) The traditional two-stage framework first adopts an instance detector to detect human and object instances and generate human-object proposals by densely connecting the detected human and object instances. Secondly, an action classification model is designed to recognize the action type for each proposal one by one. (b) We reformulate HOI detection as a point detection and matching problem in a two parallel branches framework. We represent the human/object box as the center points, widths, and heights. Based on this, we design an interaction point, *i.e.*, the midpoint of the human and object point, to link an interactive human-object pair and represent the corresponding actions. Simultaneously, we define two displacements from each interaction point to the human/object, where the human point and the object point originating from the same interaction point are considered interactive pairs. (c) We further integrate our PPDM formulation into the two-stage framework and propose a novel two-stage framework, dubbed PPDM++. PPDM++ directly detects the interactive human-object pairs firstly by PPDM, then extract region features for each human-object pair to predict the corresponding action type.

two-stage architecture heavily limits the effectiveness and efficiency of the two-stage algorithms. For effectiveness, the instance proposals are generated and filtered based on the detection confidence only while ignoring the interactiveness between a human-object instance pair, thus causing low-quality human-object proposals. Moreover, developing human-object pairwise proposals in a dense-connection manner has significantly increased the number of negative non-interactive proposals. However, an image only consists of a few interactive human-object pairs. In this case, this process not only costs a lot of computation resources unnecessarily but also increases the difficulty of finding positive samples influencing the effectiveness.

To remedy the limitation of the two-stage framework, we rethink the definition and attempt to extract key points as an intrinsical expression of instances, then represent an HOI as a point triplet <human point, interaction point, object point>. Based on this definition, we develop a top-down interaction-driven idea, where we find the interaction point first and then locate the corresponding human and object points. To this end, we reformulate HOI detection as a point detection and matching problem and design a novel one-stage *parallel* HOI detection framework, Parallel Point Detection and Matching (PPDM). As shown in Figure 2(b), our proposed PPDM breaks up the complex task of HOI detection into two simpler parallel tasks and is composed of two parallel branches. The first branch is *points detection*, which predicts HOI triplets (interaction, human and object points), and the corresponding sizes (width and height), as well as local offsets for human and object points. To predict the interaction between human and object, we select an interaction point that has the best semantic information to predict the interaction. And to match each interaction point with the human point and the object point, we design two displacements from the interaction point to its corresponding human and object point. The second branch is *points matching*, which predicts two displacements from the interaction point to its corresponding human point and object point to match each interaction point with the human point and the object point. The human and object points originating from the same interaction point are considered matched, and the three points form an HOI triplet. Our PPDM is the interaction-driven manner, where we locate interaction points first, and interaction points guide matching human-object pairs. In this way, we obtain high-quality human-object pairs considering their interactiveness and detection confidence together. 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. This saves a lot of computational costs since it avoids classifying all human-object proposals in the proposal classification stage of traditional two-stage methods.

Our novel one-stage PPDM framework offers significant improvements in both efficiency and effectiveness compared to traditional two-stage methods. However, we have identified a limitation in its interaction feature representation, which is insufficient with just a single interaction point. To address this limitation, we have integrated our novel PPDM formulation into a two-stage HOI detection framework. This integration enhances the original PPDM by providing a more comprehensive interaction feature representation while maintaining efficient interactive human-object pair detection. In traditional two-stage methods, there is a drawback of redundant non-interactive human-object proposals. However, the second stage in these methods is capable of extracting sufficient interaction features for positive human-object pairs, which enhances the representation of interactions. To overcome this limitation and focus on the essential elements, we propose a new two-stage HOI detection paradigm called PPDM++. This paradigm directly detects interactive human-object pairs in the first stage using our PPDM. Subsequently, it extracts sufficient interaction features for each human-object pair to predict the corresponding action type, similar to the second stage in traditional two-stage methods. To initiate this paradigm, we have implemented a simple framework to verify its ef-

fectiveness. We remove the classification function of PPDM and construct a Human-object Proposals Extractor (HPE) by pooling interaction and object point heatmaps from the PPDM point detect branch into a single channel. Then, we employ a simple and typical multi-branch interaction classification head to predict action and object types based on the features of the human-object region. The purpose of this simple interaction classification head is to validate the effectiveness of our PPDM-driven two-stage paradigm, and it can be easily replaced by superior interaction prediction modules in recent two-stage HOI detection methods. By adopting the PPDM++ framework, we are able to address the inherent limitation of traditional two-stage methods while retaining the advantages of two-stage frameworks. Specifically, PPDM++ allows for the extraction of representative and sufficient interaction features in a straightforward manner.

Next, we focus our attention on the HOI detection dataset. Existing datasets such as HICO-DET [37] and V-COCO [15] have made a significant contribution to the development of HOI detection. 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. HOI-A consists of $47,908$ images with 10 action types and 11 kinds of interactive objects forming 17 HOI triplets categories. The HOI-A is more application-driven and severs as a good supplement to the existing datasets. Furthermore, we organized a series of HOI detection challenges based on our HOI-A Dataset on ICCV 2019 the 2nd Person In Context workshop/challenge and CVPR 2021 the 3rd Person In Context workshop/challenge [1], which attracted more than 100 worldwide competitors.

As shown in Figure 1, experiments are conducted on the public benchmarks HICO-DET [4] and V-COCO [15], and our newly collected HOI-A dataset, and the results show that our PPDM and PPDM++ are able to achieve state-of-the-art performances in terms of accuracy and speed.

Our contributions are summarized as follows: 1) We reformulate the task of HOI detection as a point detection and matching problem and propose a novel interaction-driven one-stage PPDM framework. 2) We present the first real-time HOI detection algorithm which outperforms conventional state-of-the-art algorithms on the challenging HICO-DET, V-COCO, and HOI-A benchmarks. 3) We design a novel two-stage HOI detection pipeline based on our PPDM formulation, which directly locates interactive human-object pairs in the first stage to break the limitation from redundant non-interactive human-object proposals. 4) We build a large-scale application-oriented HOI detection dataset to supplement existing datasets.

This work is extended from our conference version [29]. We substantially revise and significantly extend the previous work in several aspects. Firstly, we propose a novel two-stage HOI detection pipeline based on our PPDM, namely PPDM++, which transforms PPDM into a human-object pair proposals extractor to directly detect interactive pairs to break the limitation from redundant human-object proposals of traditional two-stage methods. Secondly, the original PPDM is only applied to CNN-based networks as backbones. In this version, we try transformer-based backbones for PPDM and PPDM++. Thirdly, we verify the effectiveness of our proposed PPDM and PPDM++ on one more dataset, V-COCO. Fourthly, we enlarge and refine the HOI-A dataset, including more samples for each action category and accurate annotations. Finally, we elaborate more technical details and conduct more comprehensive experimental analysis, including ablation studies, quantitative comparisons, and qualitative analysis.

1. http://www.picdataset.com/

2 Related Works

2.1 Two-stage HOI Detection Methods

Conventional HOI detection methods before PPDM [29] are mostly with a two-stage framework. Firstly, such methods adopt a pre-trained object detection model [38] to detect all humans and objects in an image and then pair the detected humans and objects one by one to generate a series of human-object pairwise proposals. Secondly, a well-designed interaction prediction model is applied to predict the action categories for each detected human-object proposal. The research core of two-stage methods lies in the second stage to design a powerful interaction prediction model. Reviewing the conventional two-stage architectures, we summarize the paradigm into multi-stream and graph-based. The second stage of the multi-stream paradigm [11], [45], [44], [9], [19] mainly employs multiple convolutional streams to process cropped human and object region features, relative spatial features, or even human posture features to predict the action type for each human-object pair step-by-step. The graph-based paradigm [37], [10], [47], [42], [18], [51] first constructs a graph among a series of nodes composed of the detected human and object instances and then designs graph convolution modules to reason the interaction types between graph nodes based on the graph. Besides researches on architecture design, recent methods further explore richer features to represent HOIs, *e.g.*, linguistic features [1], [22], [53] or human structural message [44], [9], [7], [55], [16].

The above methods are all proposals based, thus their performance is limited by the quality of proposals. Additionally, the existing methods have to spend much computational cost on proposals generation and feature extraction process. To address these drawbacks, we propose a novel one-stage and interaction-driven framework to reform HOI detection task.

2.2 One-stage HOI Detection Methods

The proposal of PPDM has led to a wave of unprecedented advances for HOI detection using the one-stage pipeline. Different from two-stage methods, it outputs human-object interaction triplets directly without additional object detectors. Inspired by PPDM, point-based, box-based, and query-based methods dominate the development in the community. The point-based method [29], [54], [32]s first defines

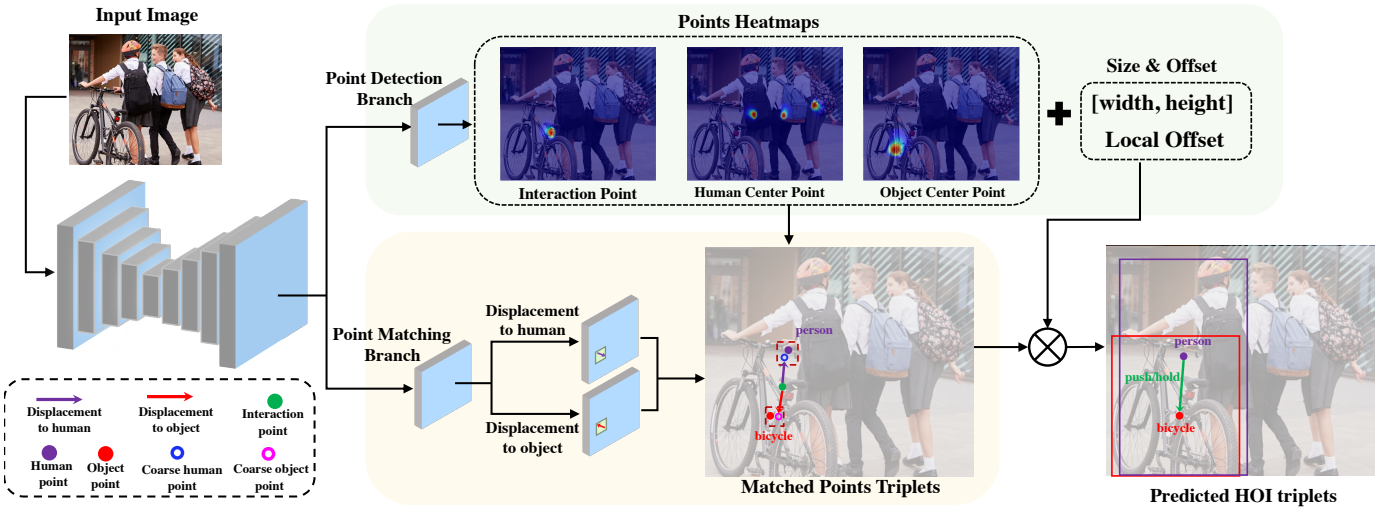

Fig. 3. **The Framework of our One-stage PPDM.** For the visual feature extraction, we adopt a dense prediction network, which produces high-resolution feature maps, to extract the appearance feature from an input image. Then we assemble two parallel branches. a) **Point Detection Branch.** With the extracted visual feature, we utilize three independent convolution blocks to estimate the heatmaps of the human/object center points and the action points. Besides, we regress the 2-D size and the local offset for each human/object point to generate the final box. b) **Point Matching Branch.** We first regress the displacements from the action point to the human point and object point, respectively, with separate convolution modules. Then, based on the estimated points and displacements, we match each action point with the corresponding human and object points to produce a set of HOI triplet results.

the key point as an interaction agent and then obtains the human-object pair through an interaction agent. The box-based methods [20], [8], [3] most adopt a union box to match human and object pair. The query-based methods recently made a giant leap with the advancement of transformer architecture [43]. It leverages self-/cross-attention to extract the global information of interaction and then yield the final triplet result through a set prediction [5], [59], [50], [21], [40].

2.3 HOI Detection Datasets.

We summarize HOI detection datasets into three categories from the label granularity: instance-level, part-level, and pixel-level. Instance-level HOI detection datasets are the most commonly used and fundamental annotation forms, which represent humans and objects as a series of bounding boxes and label the interaction categories between the human and object. There are mainly three instance-level HOI detection benchmarks: VCOCO [15], HICO-DET [4], HCVRD [58], H_2O [36]. The V-COCO is a relatively small dataset collected from the typical object detection dataset MS-COCO [31]. It selects $10,346$ images from MS-COCO and annotates with 26 actions based on the original bounding-box annotation. The HICO-DET is a large-scale HOI detection dataset for general scenes, and it has $47,776$ images with 117 verbs and 80 object categories same as COCO. The HCVRD [58] is selected from the general visual relationship detection dataset, Visual Genome [23]. It has $52,855$ images, 927 predicate categories, and $1,824$ kinds of objects. H2O is a new dataset proposed for the detection of Human-to-Human or Object Interactions (H2OIs), including both human and non-human objects. Comprising 10,301 images from the V-COCO dataset, it has been augmented with 3,666 images selected in the wild (similar to COCO dataset), mostly featuring interactions between people. Beyond traditional HOI detection datasets, the HCVRD not only focuses on human actions but also is concerned about more general human-centric relationships, *e.g.*, spatial relationships, and possessive relationships. Part-level HOI detection datasets [26] provide more fine-grain bounding-box annotations for human parts, where such datasets define and annotate a set of human part boxes, *e.g.*, head, hands, legs *et. al*. Moreover, part-level datasets provide part-object interaction for fine-grain interaction understanding. Pixel-level HOI detection datasets [33] replace bounding boxes with finer pixel-wise segmentation masks for instance-level datasets. In addition to common objects, the pixel-wise HOI detection datasets also pay attention to the interactions between humans and stuff.

In this paper, we concentrate on the instance-level HOI detection task. We review the previous HOI detection datasets and find that such datasets aim to cover all common actions in general scenes. However, for real application scenarios, we expect to be more focused. Therefore, we build up a new HOI-A dataset, which has about 41K images only annotated with limited typical kinds of actions with practical significance.

3 Parallel Point Detection and Matching

In this section, we present a detailed introduction of the pipeline of our proposed PPDM, which predicts the human-object interactive pairs and the action category simultaneously, thus generating the HOI triplets directly in a one-stage manner. In Section 3.1, we first overview the framework of PPDM. Then in Section 3.2, we introduce how to detect the human and object center points as well as the action point. Next in Section 3.3, we present how to match the estimated points to generate the HOI triplets. Finally, in Section 3.4, we show the end-to-end training process and the inference details.

3.1 Overview Framework

The HOI detection task is defined to detect the <human, object, action> HOI triplets, where human is composed of the subject bounding box and class, the object is composed of the object bounding box and class, and action indicates the interaction class. For such a complicated task, the intuitive idea is to break it up into several simpler tasks for independent optimization. Therefore, we propose the Parallel Point Detection and Matching (PPDM) framework which assembles two parallel branches for the final HOI detection results. The proposed PPDM framework is shown in Figure 3, and it includes two branches, *i.e.*, *point detection* and *point matching*. For the point detection branch, we predict each set of the corresponding center points, width and height, and local offset for representing one box candidate for both human and object. Besides, we also estimate the action point, which is defined as the midpoint of the one corresponding <human point, object point > pair. For the point matching branch, we predict the displacements between the action point and the corresponding human point and object point. Then, for the matching rule, we consider each human point and object point originated by the same action point as one matched pair.

3.2 Point Detection Branch

3.2.1 Objective Definition

The point detection branch aims to detect the human bounding box, the object bounding box as well as the action point by detecting the center points and locating the point offsets. For the human box, its center point $(x^h, y^h) \in \mathbb{R}^2$ and its width and height $(w^h, h^h) \in \mathbb{R}^2$ constitute its fundamental box info. Then, the additional point offset $\boldsymbol{\delta c^h} \in \mathbb{R}^2$ is predicted to recover discretization error caused by the output stride. For the object box, the similar strategies are implemented, and the variables are denotes as $(x^o, y^o) \in \mathbb{R}^2$, $(w^o, h^o) \in \mathbb{R}^2$ and $\boldsymbol{\delta c^o} \in \mathbb{R}^2$. Besides, we define the action point $(x^a, y^a) \in \mathbb{R}^2$ as the midpoint of a pair of human point (x^h, y^h) and object point (x^o, y^o). In this way, the receptive field of the action point can cover the interactive region of the human and object pair, thus the feature of (x^a, y^a) is feasible to estimate the action a. To be noted, the i^{th} human box among the total M human boxes are represented as $(x_i^h, y_i^h), i \in [1, M]$, and we simplify it as (x^h, y^h) if there is no confusion. The similar omission is also extended to (x^o, y^o) and (x^a, y^a).

3.2.2 Point Location

We transfer the point detection task into a heatmap estimation task following the key-point prediction method [41] by spreading the one-hot ground-truth point to a heatmap with a Gaussian kernel. As shown in Figure 3, given an image $\boldsymbol{I} \in \mathbb{R}^{3\times H\times W}$ as input, we adopt a keypoint heatmap prediction network to extract its visual feature $\boldsymbol{V} \in \mathbb{R}^{C_v \times \frac{H}{d} \times \frac{W}{d}}$, where W and H are the width and height of $\boldsymbol{I}$, and d is the output stride of the network. For the low-resolution reception field of the extracted heatmap, we calculate the corresponding low-resolution ground-truth center points. For the human point (x^h, y^h), the 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 object point is similarly defined as $(\tilde{x}^o, \tilde{y}^o) = (\lfloor \frac{x^o}{d} \rfloor, \lfloor \frac{y^o}{d} \rfloor)$. Then, we can calculate the ground-truth action point for the low-resolution heatmap by $(\tilde{x}^a, \tilde{y}^a) = (\lfloor \frac{\tilde{x}^h + \tilde{x}^o}{2} \rfloor, \lfloor \frac{\tilde{y}^h + \tilde{y}^o}{2} \rfloor)$.

In this way, we spread the three ground-truth low-resolution points $(\tilde{x}^h, \tilde{y}^h)$, $(\tilde{x}^o, \tilde{y}^o)$ and $(\tilde{x}^a, \tilde{y}^a)$ into three corresponding Gaussian heatmaps, *i.e.*, the human point heatmap $\tilde{\boldsymbol{H}}^{\boldsymbol{h}} \in [0,1]^{\frac{H}{d} \times \frac{W}{d}}$, object point heatmap $\tilde{\boldsymbol{H}}^{o} \in [0,1]^{C_o \times \frac{H}{d} \times \frac{W}{d}}$ and action point heatmap $\tilde{\boldsymbol{H}}^{a} \in [0,1]^{C_a \times \frac{H}{d} \times \frac{W}{d}}$, where C_o and C_a are the number of categories for the objects and the actions respectively. In this way, the channel number of the ground-truth object heatmap and action heatmaps are C_o and C_a respectively, and only the channel representing the specific object class or action class is non-zero. The three heatmaps are conducted by three independent convolution blocks upon the visual feature $\boldsymbol{V}$. Each convolution block is composed by a 3×3 convolution layer with ReLU, followed by a 1×1 convolution layer and a sigmoid function.

To regress the three ground-truth low-resolution heatmaps and the estimated heatmaps by the three convolution blocks, we apply an element-wise focal loss [30]. Take the action point as an example, with the estimated heatmap $\hat{\boldsymbol{H}}^{\boldsymbol{a}}$ and the ground-truth heatmap $\tilde{\boldsymbol{H}}^{\boldsymbol{a}}$, the loss is calculated as:

$$L_a = -\frac{1}{A} \sum_{k}^{C_a} \sum_{xy}^{A} \begin{cases} (1 - \hat{H}^a_{kxy})^{\alpha} \log(\hat{H}^a_{kxy}) & \tilde{H}^a_{kxy} = 1 \\ (1 - \tilde{H}^a_{kxy})^{\beta} (\hat{H}^a_{kxy})^{\alpha} & \text{else} \\ \log(1 - \hat{H}^u a_{kxy}), & \end{cases} \tag{1}$$

where A is the number of action points that equals the number of ground-truth HOI triplets in the image, and $\hat{H}^a_{kxy}$ is the heatmap score at location (x, y) for the k^{th} category in the estimated heatmaps $\hat{\boldsymbol{H}}^{\boldsymbol{a}}$. Following the default setting in [24], [57], [6], α and β are set as 2 and 4, respectively. The human point loss L_h and the object point loss L_o are calculated in the same way.

3.2.3 Size and Offset Regression

In order to obtain human and object boxes precisely, we regress the box size and the local offset from their center points. We add four convolution blocks upon the visual feature $\boldsymbol{V}$ to predict the width/height size and the local offset for the human and object boxes respectively, where each block consists of one 3×3 convolution layer with ReLU followed by one 1×1 convolution layer.

For regressing the box size and the local offset, we compute the $L1$ loss for each low-resolution ground-truth human center point $(\tilde{x}^h, \tilde{y}^h)$ and object center point $(\tilde{x}^o, \tilde{y}^o)$. The local offset loss L_{of} and the size regression loss L_s are formulary similar, and we show L_{of} as an example. We first define the ground-truth local offset for the human center point $(\tilde{x}^h, \tilde{y}^h)$ as the point location offset at the low-resolution reception field, *i.e.*, $(\tilde{\delta}^x_{(\tilde{x}^h, \tilde{y}^h)}, \tilde{\delta}^y_{(\tilde{x}^h, \tilde{y}^h)}) = (\frac{x^h}{d} - \tilde{x}^h, \frac{y^h}{d} - \tilde{y}^h)$. Then, the offset loss L_{of} is given as the average of the human box loss L^h_{of} and object box loss L^o_{of}.

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

$$L_{of}^{h} = \sum_{(\tilde{x}^h,\tilde{y}^h)\in\tilde{S}^h} (|\tilde{\delta}^{x}_{(\tilde{x}^h,\tilde{y}^h)} - \hat{\delta}^{x}_{(\tilde{x}^h,\tilde{y}^h)}| + |\tilde{\delta}^{y}_{(\tilde{x}^h,\tilde{y}^h)} - \hat{\delta}^{y}_{(\tilde{x}^h,\tilde{y}^h)}|), \tag{3}$$

$$L_{of}^{o} = \sum_{(\tilde{x}^o,\tilde{y}^o)\in\tilde{S}^o} (|\tilde{\delta}^{x}_{(\tilde{x}^o,\tilde{y}^o)} - \hat{\delta}^{x}_{(\tilde{x}^o,\tilde{y}^o)}| + |\tilde{\delta}^{y}_{(\tilde{x}^o,\tilde{y}^o)} - \hat{\delta}^{y}_{(\tilde{x}^o,\tilde{y}^o)}|), \tag{4}$$

where $\tilde{S}^h$ and $\tilde{S}^o$ denote the sets of ground-truth human points and object points respectively. M is the number of human point set $\tilde{S}^h$ and N is the number of object point set $\tilde{S}^o$. In the training dataset, M is not necessarily equal to N since one human may have interactive relationships with various objects, and form multiple HOI triplets.

3.3 Point Matching Branch

3.3.1 Objective Definition

The point matching branch aims to link the human box with the interactive object box to form one HOI triplet, and the action point is adopted as the matching bridge. In detail, the action point serves as the anchor to pair the human and object box. As shown in Figure 3, two displacements are estimated for each anchor, *i.e.*, the displacement between action point and human point $d^{ah} = (d_x^{ah}, d_y^{ah})$, and the displacement between action point and object point $d^{ao} = (d_x^{ao}, d_y^{ao})$. Then, the human point and object point can be coarsely predicted as $(x^a, y^a)+d^{ah}$ and $(x^a, y^a)+d^{ao}$ respectively. Each displacement estimation is implemented with one 3×3 convolution layer with ReLU followed by one 1×1 convolution layer. At the low-resolution reception field, the size of each displacement feature map for human and object is $2 \times \frac{H}{d} \times \frac{W}{d}$.

3.3.2 Displacement Regression

For training the point matching branch, we adopt $L1$ loss to regress the displacement of each action point. For the action point $(\tilde{x}^a, \tilde{y}^a)$, the low-resolution ground-truth displacement to the human point $(\tilde{x}^h, \tilde{y}^h)$ is computed as $(\tilde{d}^{hx}_{(\tilde{x}^a,\tilde{y}^a)}, \tilde{d}^{hy}_{(\tilde{x}^a,\tilde{y}^a)}) = (\tilde{x}^a - \tilde{x}^h, \tilde{y}^a - \tilde{y}^h)$. Similarly, the ground-truth displacement to the object point $(\tilde{x}^o, \tilde{y}^o)$ is computed as $(\tilde{d}^{ox}_{(\tilde{x}^a,\tilde{y}^a)}, \tilde{d}^{oy}_{(\tilde{x}^a,\tilde{y}^a)}) = (\tilde{x}^a - \tilde{x}^o, \tilde{y}^a - \tilde{y}^o)$. The human and object displacement predictions for $(\tilde{x}^a, \tilde{y}^a)$ are $(\hat{d}^{hx}_{(\tilde{x}^a,\tilde{y}^a)}, \hat{d}^{hy}_{(\tilde{x}^a,\tilde{y}^a)})$ and $(\hat{d}^{ox}_{(\tilde{x}^a,\tilde{y}^a)}, \hat{d}^{oy}_{(\tilde{x}^a,\tilde{y}^a)})$ respectively. Then the displacement regression loss L_{ah} and L_{ao} can be calculated as

$$L_{ah} = \frac{1}{A} \sum_{(\tilde{x}^a,\tilde{y}^a)\in\tilde{S}^a} |\hat{d}^{hx}_{(\tilde{x}^a,\tilde{y}^a)} - \tilde{d}^{hx}_{(\tilde{x}^a,\tilde{y}^a)}| + |\hat{d}^{hx}_{(\tilde{x}^a,\tilde{y}^a)} - \tilde{d}^{hy}_{(\tilde{x}^a,\tilde{y}^a)}|, \tag{5}$$

$$L_{ao} = \frac{1}{A} \sum_{(\tilde{x}^a,\tilde{y}^a)\in\tilde{S}^a} |\hat{d}^{ox}_{(\tilde{x}^a,\tilde{y}^a)} - \tilde{d}^{ox}_{(\tilde{x}^a,\tilde{y}^a)}| + |\hat{d}^{ox}_{(\tilde{x}^a,\tilde{y}^a)} - \tilde{d}^{oy}_{(\tilde{x}^a,\tilde{y}^a)}|, \tag{6}$$

where $\tilde{S}^a$ denotes the set of ground-truth action points, and A is the number of action points which is also the size of $\tilde{S}^a$.

3.3.3 Triplet matching

With the action point predictions and the human/object displacement regressions, we need to judge whether a human/object point can be matched to the action point to form an HOI triplet. We consider two conditions: 1) the human/object point needs to be close to the coarse estimated point by action point plus the corresponding displacement; 2) the human/object point needs to have a high confidence score. Based on these two considerations, for the action point $(\hat{x}^a, \hat{y}^a)$, we first rank all the detected human/object points in the point set $\hat{S}^h$ and $\hat{S}^o$, and then select the optimal points to match with the action point.

$$(\hat{x}^h_{op}, \hat{y}^h_{op}) = \underset{(\hat{x}^h,\hat{y}^h)\in\hat{S}^h}{\arg\min} \frac{1}{C^h_{(\hat{x}^h,\hat{y}^h)}} (|(\hat{x}^a, \hat{y}^a) - (\hat{d}^{hx}_{(\hat{x}^a,\hat{y}^a)}, \hat{d}^{hy}_{(\hat{x}^a,\hat{y}^a)}) - (\hat{x}^h, \hat{y}^h)|), \tag{7}$$

$$(\hat{x}^o_{op}, \hat{y}^o_{op}) = \underset{(\hat{x}^o,\hat{y}^o)\in\hat{S}^o}{\arg\min} \frac{1}{C^o_{(\hat{x}^o,\hat{y}^o)}} (|(\hat{x}^a, \hat{y}^a) - (\hat{d}^{ox}_{(\hat{x}^a,\hat{y}^a)}, \hat{d}^{oy}_{(\hat{x}^a,\hat{y}^a)}) - (\hat{x}^o, \hat{y}^o)|), \tag{8}$$

where $C^h_{(\hat{x}^h,\hat{y}^h)}$ and $C^o_{(\hat{x}^o,\hat{y}^o)}$ denote the confidence scores for the estimated human point $(\hat{x}^h, \hat{y}^h)$ and object point $(\hat{x}^o, \hat{y}^o)$ respectively.

3.4 Training and Inference

With the point location losses L_a, L_h and L_o, the displacement losses L_{ah} and L_{ao}, the size regression loss L_s and the offset loss L_{of}, the final training loss can be obtained as the weighted summation of the above losses:

$$L_{ppdm} = L_a + L_h + L_o + \lambda(L_{ah} + L_{ao} + L_s) + L_{of}, \tag{9}$$

where the weight λ is set as 0.1 following [24], [57].

The inference process is to generate the point matched HOI triplets with high confidence. Firstly, we operate a 3×3 max-pooling on the human, object, and action point heatmaps to distill the estimated points, similar to NMS operation. Secondly, we rank the point confidence scores $\hat{C}^h$, $\hat{C}^o$ and $\hat{C}^a$ to generate the top K human points $\hat{S}^h$, object points $\hat{S}^o$ and action points $\hat{S}^a$. The ranking is across the object and action categories. Next, we select matched human point and object point for each high confidence action point according to Equation 7 and 8. For a matched optimal human box $(\hat{x}^h_{op}, \hat{y}^h_{op})$, the final human box can be represented as

$$(\hat{x}^h_{rf} - \frac{\hat{w}_{(\hat{x}^h_{op},\hat{y}^h_{op})}}{2}, \hat{y}^h_{rf} - \frac{\hat{h}_{(\hat{x}^h_{op},\hat{y}^h_{op})}}{2}, \hat{x}^h_{rf} + \frac{\hat{w}_{(\hat{x}^h_{op},\hat{y}^h_{op})}}{2}, \hat{y}^h_{rf} + \frac{\hat{h}_{(\hat{x}^h_{op},\hat{y}^h_{op})}}{2}), \tag{10}$$

where $\hat{x}^h_{rf} = \hat{x}^h_{op} + \hat{\delta}^{x}_{(\hat{x}^h_{op},\hat{y}^h_{op})}$ and $\hat{y}^h_{rf} = \hat{y}^h_{op} + \hat{\delta}^{y}_{(\hat{x}^h_{op},\hat{y}^h_{op})}$ denote the human point location with the offset refinement, and the size of the corresponding box is $(\hat{w}_{(\hat{x}^h_{op},\hat{y}^h_{op})}, \hat{h}_{(\hat{x}^h_{op},\hat{y}^h_{op})})$. The object box for the matched object point $(\hat{x}^o_{op}, \hat{y}^o_{op})$ is represented in a similar manner. Finally we get the optimal set of HOI triplets with the confidence score $\hat{C}^h_{\hat{x}^h_{rf}\hat{y}^h_{rf}} \hat{C}^o_{\hat{x}^o_{rf}\hat{y}^o_{rf}} \hat{C}^a_{\hat{x}^a_{rf}\hat{y}^a_{rf}}$ for each triplet.

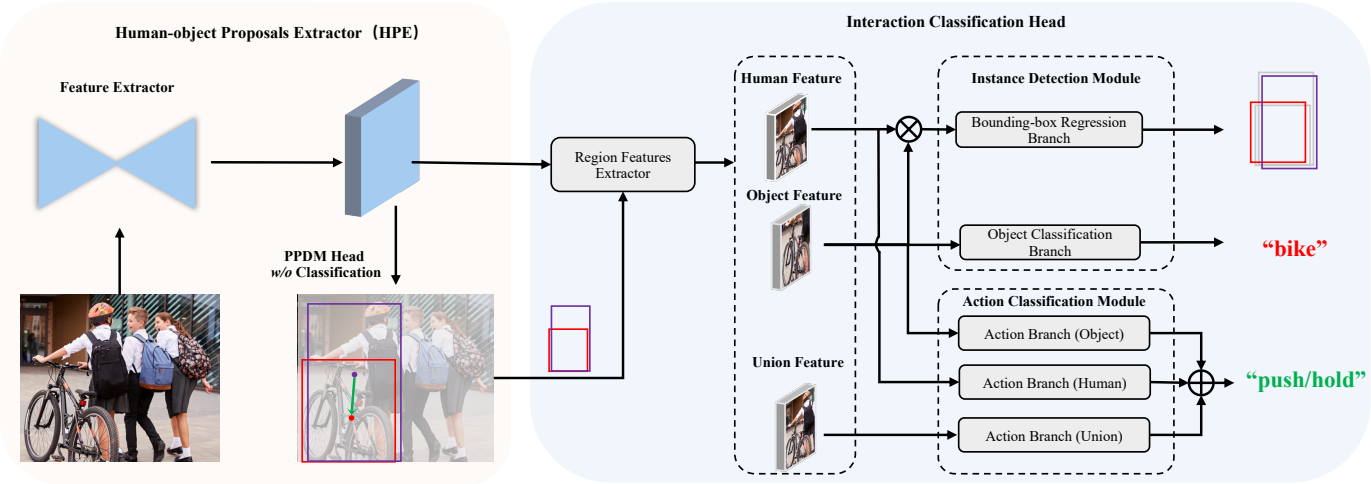

Fig. 4. **The framework of our two-stage PPDM++.** The serial pipeline consists of two modules: Human-object Proposals Extractor (HPE) and Interaction Classification Head. The former follows PPDM to detect human, object, interaction points, and the corresponding sizes and displacements, forming a series of human-object bounding-boxes pairs through the PPDM head. The latter consists of three models that first extract each human-object pair and their union region features by Region Features Extractor, then feed them into two parallel components. The instance detection module that concentrates on instances refines bounding-box localization through Bounding-box Regression Branch and obtains object type through Object Classification Branch. The action classification module leverages human-object pairs and their union region features, respectively, to predict action types. Finally, the mean function fuses the action logic, resulting in the final action type.

4 Two-stage Pipeline with PPDM

In this section, we introduce a novel end-to-end two-stage HOI detection pipeline based on our proposed PPDM, namely PPDM++. As shown in Figure 4, we first adopt PPDM to extract a series of interactive human-object pairs, then adopt an additional interaction classification head to classify action types for the detected human-object pairs. In this way, PPDM++ can not only maintain the high efficiency of PPDM by directly locating interactive pairs but also extract representative interaction features like traditional two-stage methods. In the following, we first overview the pipeline of PPDM++ in Section 4.1, then clarify how to modify PPDM into a Human-object Proposal Extractor (HPE) in Section 4.2, next elaborate the architecture of the second stage, and finally demonstrate the training and inference strategies.

4.1 The Overview of Two-stage Pipeline

In PPDM++, we break up the coupled task of HOI detection into two serial steps with an end-to-end framework. In the first step, we follow PPDM to detect human, object, and interaction points and obtain the corresponding sizes and displacements, forming a series of interactive human-object bounding-boxes pairs with type-free, <Human box, Object box>. In the second step, we first extract the human-object pair region features cropped by the human, object, and their union boxes, then we design a simple interaction classification head to predict their object and action types. Therefore, we can obtain the final HOI predictions by combining the results of the above two steps.

4.2 PPDM as Human-object Proposal Extractor

PPDM is tailored for extracting interactive human-object pairwise proposals effectively. To exploit this, we pool interaction and object point heatmaps in the PPDM point detect branch into one channel for interaction point and instance point location but remove its classification function, and the rest are remained to build the HPE. Therefore, HPE is able to concentrate on human-object pairs detection free from classification tasks. Specifically, similar to the original PPDM, we first obtain a series of human, object, and interaction points with a 2D location format, then match each interaction point with the most satisfactory human and object points through the triplet matching rule proposed in Section 3.3.3. Thus, HPE takes an image I as input and outputs a series of human-object bounding-box pairs:

$$\{[(l_i^h, t_i^h, r_i^h, b_i^h), (l_i^o, t_i^o, r_i^o, b_i^o), s_i^{ho}] | i \in [1, K^{ho}]\} = F_{hpe}(I),$$

where the first two elements in the set denote the 'ltrb' $(left, top, right, bottom)$ format human and object bounding boxes, and the third element represents the confidence score of the detected human-object pair obtained by multiplying the human object and interaction points confidence scores.

4.3 Interaction Classification Head

After human-object pairs detection, we design an interaction classification head to predict the corresponding action types and object class, and then refine human and object bounding boxes in the second stage. Given a human-object pair, we follow the typical process in general two-stage object detection methods to employ ROI align [17] to crop region features from the last-level feature of the backbone, $\boldsymbol{V}$, based on the human box, object box, and their union box, respectively:

$$\begin{aligned} \boldsymbol{V}_i^h &= F_{crop}(\boldsymbol{V}; l_i^h, t_i^h, r_i^h, b_i^h), \\ \boldsymbol{V}_i^o &= F_{crop}(\boldsymbol{V}; l_i^o, t_i^o, r_i^o, b_i^o), \\ \boldsymbol{V}_i^u &= F_{crop}(\boldsymbol{V}; \min(l_i^h, l_i^o), \min(t_i^h, t_i^o), \\ &\qquad \max(r_i^h, r_i^o), \max(b_i^h, b_i^o)), \end{aligned} \tag{11}$$

where F_{crop} denotes ROI align operation and $\boldsymbol{V}_i^h$, $\boldsymbol{V}_i^o$ and $\boldsymbol{V}_i^u$ are all with a $C_v \times H^{roi} \times W^{roi}$ sized shape.

The interaction classification head takes extracted features as inputs and adopts two parallel components, *i.e.*, instance detection module, and action classification module, to obtain final interaction triplets. Specifically, the instance detection module plays a similar role in the second stage of traditional two-stage object detectors and aims to predict object class and regress bounding-boxes offsets, $\boldsymbol{\delta}_i^{box} = (\tilde{l}_i - l_i, \tilde{t}_i - t_i, \tilde{r}_i - r_i, \tilde{b}_i - b_i)$. It consists of two branches where one branch takes $\boldsymbol{V}^o$ as input to predict the object class, and the other branch takes the concatenation of $\boldsymbol{V}^h$ and $\boldsymbol{V}^o$ as input and outputs the bounding-boxes offsets. The action classification module consists of three branches to respectively take $\boldsymbol{V}^h$, $\boldsymbol{V}^o$ and $\boldsymbol{V}^u$ as input to predict action types, and then the fused types score is obtained by a mean function for the final decision. This process can be denoted as:

$$\begin{aligned} &\boldsymbol{p}_h^a = F_h^a(\boldsymbol{V}^h), \boldsymbol{p}_o^a = F_o^a(\boldsymbol{V}^o), \boldsymbol{p}_u^a = F_u^a(\boldsymbol{V}^u), \\ &\boldsymbol{p}^a = (\boldsymbol{p}_h^a + \boldsymbol{p}_o^a + \boldsymbol{p}_u^a)/3.0; \\ &\boldsymbol{p}^o = F^o(\boldsymbol{V}^o), \boldsymbol{p}^\delta = F^\delta([\boldsymbol{V}^h, \boldsymbol{V}^o]), \end{aligned} \tag{12}$$

where $\boldsymbol{p}^o$ and $\boldsymbol{p}^a$ indicate action and object classification logits, respectively, and $\boldsymbol{p}^\delta$ denotes the predicted 4D box offsets. All branches $F^o(\cdot)$, $F^\delta(\cdot)$, $F_h^o(\cdot)$, and $F_h^a(\cdot)$ share the same architecture except output channel of last classification layer. For each branch, we first adopt a convolution to transform the ROI features into latent space, resulting in the final classification logits or offsets.

Moreover, we also explore integrating human pose and spatial features for action classification following traditional two-stage HOI detectors. Here, we simply use two additional parallel branches to process such two features and yield action logits, $\boldsymbol{p}_p^a$ and $\boldsymbol{p}_s^a$. The final form is expressed as follows:

$$\boldsymbol{p}^a = (\boldsymbol{p}_h^a + \boldsymbol{p}_o^a + \boldsymbol{p}_u^a + \boldsymbol{p}_p^a + \boldsymbol{p}_s^a)/5.0 \tag{13}$$

Besides, note that we implement a simple baseline architecture for the second stage to verify the effectiveness of our novel Two-stage PPDM pipeline. The second stage is easy to be replaced with state-of-the-art two-stage HOI detection methods to achieve advanced performance.

4.4 Training and Inference

In this subsection, we introduce the sampling mechanism and loss function during training and post-processing during inference.

The training process includes sampling strategies and optimization. During training, we first generate a series of human-object pairwise proposals through HPE following the steps in Sec. 3.4 and sample top-K_{ho} human-object pairwise proposals by confidence scores. Suppose the IoU between the human and object boxes in a human-object proposal and the corresponding boxes in any ground-truth human-object pair is greater than 0.5. In that case, we mark this proposal as a positive sample. Since a human-object pair may have more than one action, a human-object proposal is allowed to match with several ground-truth. In contrast, if the IoU is less than 0.5, the proposal is regarded as a negative sample. To ensure the training stability of the second stage, we set the ratio of positive and negative samples as 1:1 and the number of samples as K_{sample}. For the training strategies, we adopt the same optimization strategy as PPDM introduced in 3.4 and mark the loss as L_{hpe} for the first stage of PPDM++, HPE. In the second stage, we also use Focal Loss for action and object classification to keep consistency with the first stage, and L1 loss is for offset regression following traditional object detector, Faster-RCNN [38]. These losses are computed by:

$$L = \lambda_{hpe} L_{hpe} + \lambda_{oc} L_{oc} + \lambda_{ac} L_{ac} + \lambda_{reg} L_{reg} \tag{14}$$

where λ_{hpe} and L_{hpe} denotes the loss weight and loss function of the first stage of PPDM++, HPE, and the form of L_{hpe} is similar to the L_{ppdm} in Eq. 9. L_{oc}, L_{ac}, and L_{reg} represent the object classification, action classification, and regression losses, respectively.

During inference, we first follow the same process in training to sample top-K_{ho} human-object pairwise proposals by HPE. Then, we predict an action logit vector for each proposal and sample top-K_{hoi} HOI triplets across all proposals and action categories. Finally, we follow CDN [50] to conduct a pairwise NMS for the top-K_{hoi} HOI triplets to generate the final predictions.

5 HOI-A Dataset

Previous HOI datasets, *e.g.*, HICO-DET [37] and V-COCO [15], have greatly promoted the relative research of the HOI area. However, such datasets mostly concentrate on the common action categories, where some of them have less practical value. We want to pay special attention to the limited frequent but more important HOI categories, and this is not emphasized in previous datasets. Therefore, we introduce a new dataset named Human-Object Interaction for Application (HOI-A) for practical application. Our HOI-A dataset has two versions, *e.g.*, HOI-A 2019 and HOI-A 2021, where HOI-A 2021 is a refinement and extension of HOI-A 2019. In this section, we introduce the HOI-A 2021 version, while the HOI-A 2019 is introduced in our conference version [29].

We list the selected action categories and the corresponding interactive object classes of the HOI-A dataset in Table 1. All the action categories are selected driven by the significance of practical application. In this way, each action category of the HOI-A dataset can be applied in a specific scenario. For example, 'smoking' can be applied to monitoring smoking behavior at smoking-prohibited sites such as petrol stations and airplanes. Take 'talk on' as another example. We can detect some dangerous actions, *e.g.*, talking on the phone during driving which is considered dangerous driving behavior and should be reminded in some auto-driving systems. Table 2 demonstrates the detailed definition for each action category of our proposed HOI-A dataset.

5.1 Dataset Collection

In this subsection, we will present the image collection process for the HOI-A dataset from two aspects, *i.e.*, positive image collection and negative image collection.

For positive image collection, we collect images through two approaches: 1) camera shooting and 2) crawling. The

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

Verbs	Objects	# Instance
smoking	cigarette	10692
talk on	mobile phone	21667
play (mobile phone)	mobile phone	7983
eat	food	1681
drink	drink	8822
ride	bike, motorbike, horse	7429
hold	cigarette, mobile phone, food drink, document, computer	54445
kick	sports ball	600
read	document	1755
play (computer)	computer	1772

Fig. 5. 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.

camera shooting approach is important to enlarge the intra-class variation since we can design diverse camera shooting conditions. We employ 50 performers to perform the actions designed in the HOI-A dataset with different poses and take photos of them with both RGB camera and IR camera. The camera shooting procedure is executed under various illumination and scene conditions. The crawling approach is another important way to complement positive samples. To crawl data from the internet, we generate a series of keywords including the HOI triplet phrase following such format 'a person [Verb-ing] a/an [Object]', the action pair with the format '[Verb-ing] a/an [Object]', and the action names. The retrieved images from the internet are then cleaned and prepared for annotation.

For negative image collection, we define two kinds of negative samples for each set of <human, action, object > triplet. The first situation is that the image has the concerned object but has no interactive actions. For example, in Figure 5(f), the man is not smoking a cigarette although the cigarette exists in the image. In this case, the image is collected as a negative sample. The second situation is that the concerned action happens, but no interactive objects exist in the image. For example, Figure 5(e) seems to be a man smoking at a glance, however, the image actually has no cigarette with a closer look. For this situation, we design an 'attack' manner to collect the hard negative samples. We first train a multi-label action classifier with the annotated positive samples, which outputs the confidence of each action category of the input image. Then, we require the performers to arbitrarily attack the classifier with the pre-defined action but no interactive objects. Finally, we record the photo of a successful attacking case to be one hard negative sample.

5.2 Dataset Annotaion

The HOI-A dataset annotation process includes box annotation and action annotation. Firstly, we annotate all the objects of pre-defined classes with a bounding box and the corresponding category. To be noted, the human is regarded as a specific category of all the objects during the box annotation process. Secondly, we show the image visualization with the boxes and corresponding IDs, and annotate whether a human box ID is interactive with an object box ID. Therefore, we record the interactive <human ID, action ID, object ID> set as one annotated triplet. To guarantee annotation accuracy, 3 annotators annotate one image repeatedly, and only the triplets annotated by more than 2 annotators are regarded as qualified annotations.

TABLE 2
The definitions of the verbs in HOI-A dataset.

Verbs	Definitions
smoking	1.The cigarette is in the mouth; 2. The cigarette in the hand with smoke around it. 3. The burning cigarette in the hand (part of the cigarette is soot)
talk on	The mobile phone is held in hand near the ear.
play (mobile phone)	The mobile phone is held in hand away from ear, and the person looks at the mobile phone.
eat	1. The food is held near the mouth and the person is eating; 2. The food in the hand (probably away from the mouth), but there is an obvious act of chewing.
drink	The water cup is near the mouth and the person is drinking.
ride	The person rides on horses, bicycles, electric bicycles or motorcycles.
hold	Smoking, call, drink, etc., as long as there is something held in the hand(s).
kick	One must take the action of kicking. (If it is only the ball next to the feet, but without the action of kicking, it is not kicking.)
read	Including newspapers, books, and other paper materials, requiring eyes to look at the materials
play (computer)	including the use of desktops, laptops, etc.

5.3 Dataset Properties

Variation. As we mentioned in Section 5.1, we capture photos with different scenes and conditions in the image collection process to enlarge the intra-class variation. In detail, for each type of HOI triplet in the HOI-A dataset, we take photos with 3 general scenes (*i.e.*, indoor, outdoor, and in-car), 3 lighting conditions (*i.e.*, dark, natural and intense), various human poses and different angles. Besides, we shoot each of the photos with both RGB and IR cameras.

Scale. The HOI-A dataset contains $47,908$ annotated images with 11 object categories and 10 action categories, which

TABLE 3
Performance comparison on the HICO-DET test set. We use 'A', 'P', 'L' to denote the appearance feature, human pose information, and the language feature for interaction classification, respectively.

Method	Feature	Default			Know Object			Inference Time (ms) ↓	FPS ↑
		Full	Rare	Non-Rare	Full	Rare	Non-Rare		
DETR-based Methods:									
HOI-trans [59]	A	23.46	16.91	25.41	19.24	19.24	28.22	68	14.71
HOTR [21]	A	25.10	17.34	27.42	-	-	-	75	13.33
AS-Net [5]	A	28.87	24.25	30.25	31.74	27.07	33.14	75	13.33
QPIC-R50 [40]	A	29.07	21.85	31.23	31.68	24.14	33.93	68	14.71
CDN-R50 [50]	A	31.44	27.39	32.64	34.09	29.63	35.42	68	14.71
UPT [52]	A	31.66	25.94	33.36	35.05	29.27	36.77	65+61=126	7.94
RCNN-based Methods:									
Shen *et. al* [39]	A + P	6.46	4.24	7.12	-	-	-	-	-
HO-RCNN [4]	A	7.81	5.37	8.54	10.41	8.94	10.85	-	-
InteractNet [13]	A	9.94	7.16	10.77	-	-	-	145	6.90
GPNN [37]	A	13.11	9.34	14.23	-	-	-	197 + 48 = 245	4.08
Xu *et. al* [46]	A + L	14.70	13.26	15.13	-	-	-	-	-
iCAN [11]	A	14.84	10.45	16.15	16.26	11.33	17.73	92 + 112 = 204	4.90
PMFNet-Base [44]	A	14.92	11.42	15.96	18.83	15.30	19.89	-	-
Wang *et. al* [45]	A	16.24	11.16	17.75	17.73	12.78	19.21	-	-
No-Frills [16]	A + P	17.18	12.17	18.68	-	-	-	197 + 230 + 67 = 494	2.02
TIN [27]	A + P	17.22	13.51	18.32	19.38	15.38	20.57	92 + 98 + 323 = 513	1.95
RPNN [55]	A + P	17.35	12.78	18.71	-	-	-	-	-
PMFNet [44]	A + P	17.46	15.65	18.00	20.34	17.47	21.20	92 + 98 + 63 = 253	3.95
Our Point-based Methods:									
PPDM-DLA	A	19.94	13.01	22.01	22.63	15.93	24.63	**24**	**41.67**
PPDM-Hourglass	A	21.73	13.78	24.10	24.58	16.65	26.84	71	14.08
PPDM-SwinT	A	22.26	12.64	25.13	24.19	14.24	27.17	40	25
PPDM-SwinB	A	27.59	18.07	30.44	29.11	18.85	32.71	58	17.54
PPDM++-DLA	A	23.34	16.64	26.34	26.22	19.46	28.23	47	21.28
PPDM++-Hourglass	A	25.82	18.54	27.99	28.53	20.90	30.81	90	11.11
PPDM++-SwinT	A	25.49	18.77	27.49	28.13	21.28	30.17	59	16.95
PPDM++-SwinB	A	**30.10**	**23.73**	**32.00**	**31.80**	**24.93**	**33.85**	79	12.66
+ 1 Transformer Encoder Layer	A	30.84	23.96	32.90	32.44	24.98	34.67	83	12.05
+ Multi-level ROI Features	A	31.20	25.02	33.05	32.85	26.18	34.84	97	10.31

forms 17 kinds of HOI combinations. In detail, HOI-A has $52,904$ human instances, $70,951$ object instances and $116,846$ action instances. Each human interacts with 2.2 objects on average. We present the number of instances for each verb in Table 1. The instance number is at least 360 while 60% of the verbs appear more than $6,500$ times in the dataset. To the best of our knowledge, HOI-A is already the largest existing HOI dataset in terms of the image number per action category. In HOI-A, the 'hold' is the most frequent verb because 'hold' often appears simultaneously with other verbs. For example, if someone is playing mobile phone, he should also hold the phone. For the experiments, we split the dataset into two parts, $38,067$ images for training and $9,841$ images for testing. For fair evaluation, we keep the same ratios of each verb in the training set and test set.

6 Experiments

In this section, we introduce the experimental details, where experimental settings in Section 6.1, the comparisons with traditional methods in Section 6.2, ablation studies in Section 6.3, and detailed qualitative analysis in Section 6.4.

6.1 Experimental Setting

Datasets. Experiments are conducted on three HOI detection benchmarks, including HICO-DET [4], V-COCO, and proposed HOI-A datasets, to prove the effectiveness of proposed PPDM and PPDM++. HICO-DET consists of $47,776$ images ($38,118$ for training and $9,658$ for testing). It has 600 HOI categories in the form of <action, object> over 117 action categories, and 80 object categories, same as MS-COCO dataset [31]. 138 types of them are considered as the rare HOIs which appear less than 10 times, and 462 kinds of HOIs are regarded as the non-rare set. V-COCO is the subset of MS-COCO, which consists of $5,400$ images in the 'trainval' dataset and $4,946$ images in the test set. Each human is annotated with binary labels for 29 different action categories. We transform the annotations of V-COCO, HICO-DET, and HOI-A datasets into a unified form to facilitate the experiments. For the annotation of each image, we first include all human and object bounding boxes in a list, then represent an HOI as <h-id, o-id, v-id>, where h-id and o-id denote the index of the corresponding human and object boxes in the list, and v-id is the action categories index.

Metric. Following the standard scheme, we use mean average precious (mAP) to examine the model for three benchmarks. A predicted triplet is considered a true positive sample if it detects the human and object accurately, *i.e.*, the Interaction-over-Union (IOU) between the detected bounding-box and ground-truth is large than 0.5 and predicts the correct action category. Specially, we compute AP per HOI class in HICO-DET and V-COCO and compute AP per verb class in the HOI-A dataset.

Implementation Details. We adopt two dense prediction CNN-based backbones, Hourglass-104 [35], [24] and DLA-

34 [49], [57], and a popular transformer-based backbone, Swin-Transformer [34], as the visual feature extractor of PPDM and PPDM++ to conduct our experiments. For the CNN backbones, we follow CenterNet [57] to use the modified versions for point prediction tasks. We initialize the feature extractor with the weights from CenterNet pre-trained on MS-COCO [31]. CenterNet [57] only provided the pre-trained weights for the CNN-based backbones, so we construct a Swin-Transformer-based CenterNet and trained it on the MS-COCO dataset. We adopt an FPN-style architecture for Swin-Transformer, and its performances on MS-COCO are 38.4 and 42.5 for SwinT and SwinB, respectively. Here, for Swin-Transformer, we pre-train it on the MS-COCO dataset with 140 epochs. All experiments are implemented with PyTorch 1.5 and CUDA 10.0. We train CNN-based backbone for PPDM and PPDM++ with Adam on 8 1080Ti GPUS, while transformer-based architectures are trained with AdamW with 4 A100 GPUs.

Next, we introduce parameter settings during PPDM and PPDM++ training and inference. We follow CenterNet [57] to transform the input image as fixed size 512×512 with 'random scale and random shift then crop' data augmentation, and the corresponding output feature size is 128×128 for all backbones. We train all backbones for 110 epochs and step the learning rate at the 90th epoch by 10 times. The batch size for DLA-34 is 64, and the initialization learning rate is $3e$-4. We train the hourglass-based model with a 16 sized mini-batch with a $1e$-4 initialization learning rate. For Swin-based models, we adopt a $3e$-5 learning rate with a 60 batch size. We set the number of selected predictions K as 100. For the loss weights setting, we follow CenterNet to set the first-stage PPDM loss weights, and the interaction and box classification loss weights are 10 and the regression loss weight is 1 in the second stage.

6.2 Comparison to State-of-the-art

We conduct experiments on three HOI detection benchmarks and compare our proposed PPDM and PPDM++ with previous state-of-the-art HOI detection methods from the performance and efficiency views (HICO-DET in Table 3, V-COCO [15] in Table 4 and HOI-A in Table 5). Note that we only compare the methods published earlier than PPDM [28] since recent methods [40], [5], [50], [59] benefit from the superior query mechanism in the transformer detection framework, *e.g.*, DETR [2], to achieve strong performances. However, our PPDM and PPDM++ adopt the dense prediction detection pipeline. Moreover, such methods are also inspired by the PPDM proposed one-stage framework. Next, we show a detailed comparison and analysis of the three benchmarks, respectively.

HICO-DET. As shown in Table 3, comparing conventional instance-driven two-stage methods, our novel interaction-driven one-stage framework PPDM has significantly outperformed such methods across all backbones, though previous methods employ additional features. Especially when equipped with the advanced transformer-based backbone, Swin-transformer, our PPDM is able to achieve 27.59 mAP only with appearance features. Furthermore, benefiting from sufficient interaction feature representation, our proposed PPDM++ further improves the performance, where PPDM++ with SwinB backbone has achieved 30.10 mAP. In the same backbone setting, the performance improvement from PPDM to PPDM++ is about 2.5 mAP, which is a significant gain proving the strong interaction feature representation of PPDM++.

TABLE 4
Performance comparisons on V-COCO test set. We mark the appearance feature, human pose feature, and language feature as 'A', 'P', and 'L', respectively.

Methods	Feature	AP_{role}
DETR-based Methods:		
HOI-trans [59]	A	52.9
AS-Net [5]	A	53.9
HOTR [21]	A	55.2
QPIC [40]	A	58.3
CNN-based Methods:		
Gupta *et. al* [15]	A	31.8
InteractNet [13]	A	40.0
RPNN [55]	A+P	47.5
UnionDet [22]	A	47.5
TIN (RP_DC_D) [27]	A+P	47.8
VCL [19]	A	48.3
C-HOI [56]	A	48.3
DRG [10]	A+L	51.0
VSGNet [42]	A+S	51.7
PMFNet [44]	A+P	52.0
PPDM-Hourglass [29]	A	50.9
PPDM++-Hourglass [29]	A	54.3

V-COCO. See Table 4, we conduct experiments based on the Hourglass backbone in the V-COCO dataset. Our PPDM and PPDM++ have achieved comparable and superior performance comparing traditional two-stage methods, where our PPDM++ has outperformed the previous state-of-the-art two-stage method, PMFNet, 2.3 point. This is primarily attributed to the inherent characteristics of the V-COCO dataset, including its scale and origin (derived from the MS-COCO dataset), which substantially influences performance metrics, heavily reliant on the choice of pretrained detection models. Compared to methods based on DETR, our approach does not exhibit superiority. Notably, DETR demonstrates a more robust performance compared to CenterNet-Hourglass.

TABLE 5
Performance comparison on HOI-A 2019 and 2021 test sets.

Method	Dataset	mAP (%)	Time (ms)
Faster Interaction Net	HOI-A 2019	56.93	-
GMVM	HOI-A	60.26	-
C-HOI [56]	HOI-A 2019	66.04	-
iCAN [11]	HOI-A 2019	44.23	194
TIN [27]	HOI-A 2019	48.64	501
PPDM-DLA	HOI-A 2019	67.45	**27**
PPDM-Hourglass	HOI-A 2019	**71.23**	71
QPIC-R50 [40]	HOI-A 2021	77.57	67
CDN-R50 [50]	HOI-A 2021	78.25	67
PPDM-Hourglass	HOI-A 2021	76.23	71
PPDM++-Hourglass	HOI-A 2021	**79.12**	89

HOI-A. Since HOI-A is a newly collected dataset, there are no existing results from conventional methods. Aiming to build up a benchmark and conduct a sufficient comparison, we select two typical open-source HOI detection methods and the top-3 results from the leaderboard of the ICCV

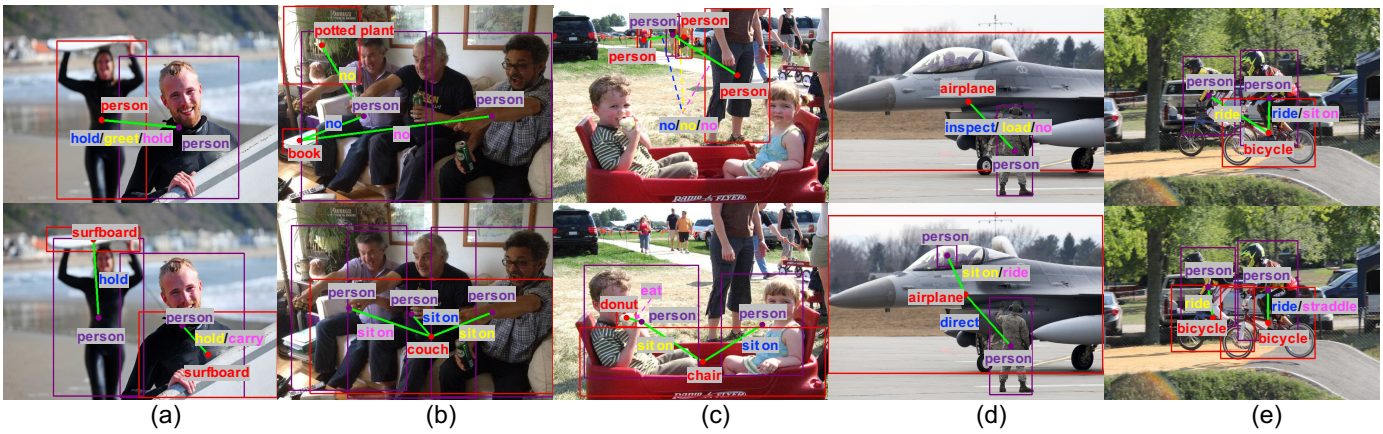

Fig. 6. Qualitative comparison between iCAN (top) and our PPDM (bottom) through HOI prediction visualization on HICO-DET. We use 'Purple' and 'Red' to represent humans and objects, respectively, and link the interactive human and object pairs with a green line. For a clear vision, we only keep the three most confident predictions for visualization, where the top 3 predictions are separated by different colors (1-blue, 2-yellow, 3-pink).

2019 PIC challenge HOI detection track [2]. Additionally, we compare our PPDM and PPDM++ to the baseline methods in two versions of HOI-A, *i.e.*, HOI-A 2019 and HOI-A 2021. Our PPDM has outperformed the winning method, C-HOI [56], though C-HOI adopts a powerful instance detector to produce proposals. Additionally, we choose iCAN [11] and TIN [27] to construct the baselines on HOI-A 2019 dataset. We follow the original settings in such methods to pre-train an instance detector, *i.e.*, Faster-RCNN, on the HOI-A dataset and train their corresponding HOI classifiers. Comparing such two-stage methods, our PPDM has achieved a significant improvement. Additionally, we further conduct experiments on HOI-A 2021 dataset to verify the effectiveness of our PPDM and PPDM++. In such an extension and clean version, our PPDM is able to achieve more satisfactory performance. Comparing recent DETR-based methods, QPIC [40] and CDN [50], our PPDM++ has also achieved superior performance. Thus, our PPDM and PPDM++ can achieve a very high absolute performance in such practical significant HOI label space, significantly pushing the HOI toward practical application.

Efficiency Analysis. We conduct a detailed efficiency comparison and analysis on a single RTX 1080Ti GPU. Note that we only report the inference speed of open-source methods, and the speed computation includes all steps from an image to produce final HOI predictions, *i.e.*, instance detection, HOI classification, and human pose estimation. Taking HICO-DET as the exemplar dataset, PPDM-DLA is the first real-time HOI detection method with 42 fps, which is about 10 times faster than the state-of-the-art two-stage method, PMFNet, due to the superior parallel bottom-up framework. Though the second stage increases the computation cost, PPDM++ is still significantly faster than the previous two-stage methods. Especially, PPDM++ with a heavy backbone, SwinB, is also 3 times faster than PMFNet. The inference time of the same methods in the HOI-A dataset is less than those in the HICO-DET dataset since the number of HOI categories in the HOI-A dataset is fewer than in the HICO-DET dataset leading to fewer parameters and less computation.

Comparion with DETR-based Methods. We select several representative DETR-based HOI detection methods for comparison. Comparing recent SOTA DETR-based methods, our PPDM or PPDM++ remains a performance gap in the same backbone setting. The gap between our PPDM or PPDM++ between recent SOTA methods lies in the different detection frameworks, *i.e.*, point-based *v.s.* query-based, where DETR [2] is a stronger object detection paradigm than CenterNet [57]. Though equipped with a strong backbone SwinB, CenterNet has achieved similar object detection performance with DETR-R50 (42.5% *v.s.* 42.0%). Under the 'similar detection performance' setting, our PPDM++ with SwinB (30.10%) has achieved comparable or superior performance with the typical DETR-based HOI detection methods, *e.g.*, QPIC [40] (29.70%), HOTR [21] (25.10%), HOI-trans [59] (26.61%), AS-Net [5] (28.87%) in HICO-Det dataset. It proves that our PPDM++ is a competitive HOI detection paradigm. In the efficiency comparison, the inference time of single-branch DETR-based HOI detection methods QPIC and HOI-trans in a single Nvidia 1080Ti is $68ms$, and of the two-branch methods HOTR and AS-Net is $75ms$. Our PPDM++-SwinB is able to achieve comparable inference speed with these methods, though with a large-scale backbone.

Moreover, the transformer encoder in DETR can provide rich global context information, which is beneficial for interaction understanding since complex interaction understanding requires querying useful features from the whole image, *e.g.*, human pose, interactive region, scene context, and spatial location information. We have conducted experiments by adding a single transformer encoder layer followed by the Swin-B backbone. With this modification, though the object detection performance of CenterNet has not improved (42.1%), the HOI detection performance has improved 0.74% (30.1%-30.84%). Additionally, we have conducted experiments based on the 'extracting human/object/union region features from multi-resolution backbone feature maps' setting on the HICO-Det dataset, where we have achieved 31.20% mAP.

Compared to recent SOTA DETR-based HOI detection

2. http://www.picdataset.com/challenge/leaderboard/hoi2019

methods, their performance improvement mainly comes from complex and superior interaction prediction modules (UPT [52]) or extra fine-grain human part annotations or large-scale pre-trained models. However, PPDM++ is a novel paradigm, and its performance is also able to improve by refining the second stage with recent complex and superior interaction prediction modules.

6.3 Ablation Study

In this subsection, we conduct ablation studies to analyze the effectiveness of components in PPDM and PPDM++.

TABLE 6
Ablation studies on HICO-DET Test Set.

Method	Full	Rare	Non-Rare	Time
Union Center	18.65	12.11	20.61	24
PPDM-DLA	19.94	13.01	22.01	24
+ UA branch	21.65	14.47	23.79	43
+ UA&UH branch	22.05	14.54	24.29	45
+ UA&UH&UO branch	22.65	14.89	24.97	45
PPDM++-DLA	23.34	16.64	26.34	47
- PNMS	22.46	16.23	24.32	42
- Reg&PNMS	22.37	16.06	24.23	42

Backbones. In Table 3, we explore the effectiveness of our PPDM and PPDM++ with various visual backbones, *i.e.*, DLA-34, Hourglass-104, SwinT, and SwinB, and conduct experiments and evaluation on the HICO-DET dataset. Here, we conduct a comparison with the CNN-based backbone and transformer-based backbone. CenterNet equipped with DLA-34 or SwinT can achieve very similar performance in the MS-COCO dataset, while in HOI detection, PPDM/PPDM++ with SwinT outperforms it with DLA-34 a large margin, about 2 mAP. Moreover, with SwinB backbone, PPDM and PPDM++ can achieve powerful performance while keeping a faster speed than Hourglass-104. We conclude that transformer-based backbones can provide a larger receptive field and global context, which are beneficial to interaction understanding.

Interaction Point Selection. We further explore the location of interaction points selection and verify the reasonability of the midpoint choice. To this end, we select another reasonable-sound point, the center of the union of human and object bounding-boxes, as the interaction point to perform an experiment. See the 'Union Center' setting in Table 6. With this interaction point, the mAPs drop 1.29 points compared with the original midpoint setting. We attribute this performance drop that two objects may often interact with the same human while in the human box, thus causing complete overlap between the center points of their union boxes. In this case, the union center setting fails to detect such two HOI triplets at the same time.

Interaction Head Setting. Here, we conduct a series of experiments to analyze the effectiveness of different branches for interaction classification. As shown in Table 6, we first only adopt union features ('UA branch') to predict actions and achieve 21.65 mAP. When adding human region features ('UH branch') to assist action prediction, the performance has improved 0.4 mAP. And fusing object region features has further boosted the performance. Finally, we integrate the relative spatial features into the union branch, which has achieved 0.59 performance gain. Thus, employing diversity interaction feature representation can produce accurate HOI predictions. However, we only attempt a simple interaction head architecture to verify the effectiveness of our PPDM++ framework. Adapting superior HOI classification mechanisms in two-stage methods can further improve PPDM++'s performance.

Post-Processing. Here, we analyze the post-processing operations in our PPDM++ during inference. Firstly, we produce the HOI predictions without using PNMS, which causes a 0.88 mAP drop. Secondly, we remove the bounding-box refinement operation by the second-stage regression branch. In this case, the performance has dropped 0.1 mAP.

Strong Detector. We have implemented a more stronger detection framework, employing DeformableDETR-SwinL, fine-tuned on the HICO-DET dataset, for generating human and object bounding boxes. This approach supersedes our previous method of predicting boxes. Specifically, for the task of point triplets matching, we derive human and object points from these enhanced bounding boxes, which are then utilized to form the final Human-Object Interaction (HOI) triplets. The integration of these superior boxes has notably improved our model's performance. Concretely, our modified PPDM++-Swin-B model now attains a mAP of 33.4 on the HICO-DET dataset, marking an improvement of approximately 3 mAP points. Additionally, when applying the DeformableDETR-SwinL detector, trained on the MS-COCO dataset, for the V-COCO dataset, our PPDM++-Hourglass model achieves an AP of 56.76. These enhancements clearly demonstrate the significant benefits of employing advanced detection mechanisms in augmenting HOI detection capabilities.

6.4 Qualitative Analysis

In this subsection, we present a comparison and analysis from a qualitative view, thoroughly.

Prediction Visualization and Comparison. Here, we conduct a detailed comparison of the traditional two-stage framework with our one-stage PPDM framework in a result visualization manner. For this goal, we choose the representative two-stage method iCAN [11] for comparison. We summarize several general bad cases in the two-stage framework and show the visualization results in Figure 6, where we select and visualize three predicted HOI results with the highest confidence scores for each image. Due to the serial two-stage framework, the most common case is that traditional two-stage easily gives high confidence for a non-interactive instance, which has a high instance detection confidence. As shown in Figure 6(b) and Figure 6(c), since iCAN suffers from large-scale negative samples generated in the first stage, it tends to predict a series of high-confidence HOI triplets of 'non-interaction' type. In contrast, our PPDM can significantly alleviate such problems from a bottom-up interaction-driven pipeline. See Figure 6(d). Though the pilot in the airplane is so small and heavily occluded and hard to be detected in an instance detector, our PPDM can accurately predict the HOI triplets with high confidence in these cases. We attribute it that PPDM is interactive-driven and free from pre-predicted proposals. Therefore, PPDM concentrates on the HOI understanding and is a superior HOI detection pipeline.

Fig. 7. We visualize the interaction points and the corresponding displacements, where the Red and purple lines respectively denote displacements from the interaction point (green) to humans and objects.

Element Visualization. Here, we show the detailed visualizations for predicted elements in our PPDM, *i.e.*, points and displacements. As shown in Figure 7, our PPDM can accurately predict interaction points located at the midpoint of the corresponding human and object center points, though the human is far away from the object or lies in the object region. For further displacement understanding, we also show the displacements in Figure 7. PPDM is able to produce accurate displacement predictions where the interaction point plus the displacement is very close to the corresponding instance center point.

7 Conclusion

In this paper, we propose a novel formulation for the HOI detection problem, where we break the traditional two-stage top-down instance-driven framework into a novel one-stage bottom-up interaction-driven framework, PPDM. We define an HOI triplet as a point triplet and adopt the interaction as the midpoint of the corresponding human and object center points, and design a parallel framework. In this way, we locate interaction first and then find the corresponding interactive instances. In this way, our PPDM can easily concentrate on the interactive regions and produce accurate HOI predictions. Moreover, our PPDM is the first real-time HOI detection method since it is free from traditional serial searching HOI classification manner. Additionally, we integrate our novel PPDM formulation into a two-stage pipeline, PPDM++, for sufficient interaction representation. Though our PPDM++ is a two-stage method, it is also an interaction-driven pipeline, which saves a lot compuation cost. Our PPDM++ futher improve the performance of PPDM. Finally, we build up an newly HOI detection benchmarks for practical application, namely, HOI-A.

References

[1] Ankan Bansal, Sai Saketh Rambhatla, Abhinav Shrivastava, and Rama Chellappa. Detecting human-object interactions via functional generalization. In *AAAI*, 2020.

[2] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, 2020.

[3] Sanaa Chafik, Astrid Orcesi, Romaric Audigier, and Bertrand Luvison. Classifying all interacting pairs in a single shot. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2892–2901, 2020.

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

[5] Mingfei Chen, Yue Liao, Si Liu, Zhiyuan Chen, Fei Wang, and Chen Qian. Reformulating hoi detection as adaptive set prediction. In *CVPR*, 2021.

[6] 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.

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

[8] Hao-Shu Fang, Yichen Xie, Dian Shao, and Cewu Lu. Dirv: Dense interaction region voting for end-to-end human-object interaction detection. In *AAAI*, 2021.

[9] 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. In *AAAI*, 2019.

[10] Chen Gao, Jiarui Xu, Yuliang Zou, and Jia-Bin Huang. Drg: Dual relation graph for human-object interaction detection. In *ECCV*, 2020.

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

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

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

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

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

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

[17] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, pages 2961–2969, 2017.

[18] Tao He, Lianli Gao, Jingkuan Song, and Yuan-Fang Li. Exploiting scene graphs for human-object interaction detection. In *ICCV*, 2021.

[19] Zhi Hou, Xiaojiang Peng, Yu Qiao, and Dacheng Tao. Visual compositional learning for human-object interaction detection. In *ECCV*, 2020.

[20] Bumsoo Kim, Taeho Choi, Jaewoo Kang, and Hyunwoo J. Kim. Uniondet: Union-level detector towards real-time human-object interaction detection. In *ECCV*, 2020.

[21] Bumsoo Kim, Junhyun Lee, Jaewoo Kang, Eun-Sol Kim, and Hyunwoo J. Kim. Hotr: End-to-end human-object interaction detection with transformers. In *CVPR*, 2021.

[22] Dong-Jin Kim, Xiao Sun, Jinsoo Choi, Stephen Lin, and In So Kweon. Detecting human-object interactions with action co-occurrence priors. In *ECCV*, 2020.

[23] 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.

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

[25] Yong-Lu Li, Xinpeng Liu, Xiaoqian Wu, Xijie Huang, Liang Xu, and Cewu Lu. Transferable interactiveness knowledge for human-object interaction detection. *TPAMI*, 2021.

[26] Yong-Lu Li, Xinpeng Liu, Xiaoqian Wu, Yizhuo Li, Zuoyu Qiu, Liang Xu, Yue Xu, Hao-Shu Fang, and Cewu Lu. Hake: A knowledge engine foundation for human activity understanding, 2022.

[27] 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.

[28] Yue Liao, Si Liu, Guanbin Li, Fei Wang, Yanjie Chen, Chen Qian, and Bo Li. A real-time cross-modality correlation filtering method for referring expression comprehension. In *CVPR*, 2020.

[29] Yue Liao, Si Liu, Fei Wang, Yanjie Chen, Chen Qian, and Jiashi Feng. Ppdm: Parallel point detection and matching for real-time human-object interaction detection. In *CVPR*, 2020.

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

[31] 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.

[32] Xue Lin, Qi Zou, and Xixia Xu. Action-guided attention mining and relation reasoning network for human-object interaction detection. In *IJCAI*, 2020.

[33] Si Liu, Zitian Wang, Yulu Gao, Lejian Ren, Yue Liao, Guanghui Ren, Bo Li, and Shuicheng Yan. Human-centric relation segmentation: Dataset and solution. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[34] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

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

[36] Astrid Orcesi, Romaric Audigier, Fritz Poka Toukam, and Bertrand Luvison. Detecting human-to-human-or-object (h 2 o) interactions with diabolo. In *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*, pages 1–8. IEEE, 2021.

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

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

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

[40] Masato Tamura, Hiroki Ohashi, and Tomoaki Yoshinaga. Qpic: Query-based pairwise human-object interaction detection with image-wide contextual information. In *CVPR*, 2021.

[41] Jonathan J Tompson, Arjun Jain, Yann LeCun, and Christoph Bregler. Joint training of a convolutional network and a graphical model for human pose estimation. In *NIPS*, 2014.

[42] Oytun Ulutan, ASM Iftekhar, and Bangalore S Manjunath. Vsgnet: Spatial attention network for detecting human object interactions using graph convolutions. In *CVPR*, 2020.

[43] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, 2017.

[44] 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.

[45] 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.

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

[47] Dongming Yang and Yuexian Zou. A graph-based interactive reasoning for human-object interaction detection. In *IJCAI*, 2020.

[48] 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.

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

[50] Aixi Zhang, Yue Liao, Si Liu, Miao Lu, Yongliang Wang, Chen Gao, and Xiaobo Li. Mining the benefits of two-stage and one-stage hoi detection. In *NIPS*, 2021.

[51] Frederic Z Zhang, Dylan Campbell, and Stephen Gould. Spatially conditioned graphs for detecting human-object interactions. In *ICCV*, 2021.

[52] Frederic Z Zhang, Dylan Campbell, and Stephen Gould. Efficient two-stage detection of human-object interactions with a novel unary-pairwise transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20104–20112, 2022.

[53] Xubin Zhong, Changxing Ding, Xian Qu, and Dacheng Tao. Polysemy deciphering network for robust human–object interaction detection. *IJCV*, 2021.

[54] Xubin Zhong, Xian Qu, Changxing Ding, and Dacheng Tao. Glance and gaze: Inferring action-aware points for one-stage human-object interaction detection. In *CVPR*, 2021.

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

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

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

[58] Bohan Zhuang, Qi Wu, Chunhua Shen, Ian Reid, and Anton van den Hengel. Hcvrd: A benchmark for large-scale human-centered visual relationship detection. In *AAAI*, 2018.

[59] Cheng Zou, Bohan Wang, Yue Hu, Junqi Liu, Qian Wu, Yu Zhao, Boxun Li, Chenguang Zhang, Chi Zhang, Yichen Wei, and Jian Sun. End-to-end human object interaction detection with hoi transformer. In *CVPR*, 2021.

Yue Liao is currently a Ph.D. candidate at School of Computer Science and Engineering, Beihang University. He received his Master degree from Institute of Information Engineering, Chinese Academy of Sciences. His research interests include human-object interaction detection, visual grounding and object detection. He has published more than 10 papers at top journals and conferences, including T-PAMI, T-IP, NIPS, CVPR and ECCV, etc. He was the Champion of CVPR 2021 ActivityNet Homage Challenge. He has been serving as a reviewer for numerous academic journals and conferences, such as TCSVT, TMM, CVPR, ICCV, ECCV, AAAI and ACM MM.

Si Liu is currently a professor at Beihang University. She received her Ph.D. degree from Institute of Automation, Chinese Academy of Sciences. She has been a research assistant and Postdoc in National University of Singapore. Her research interest includes computer vision and multimedia analysis. She has published over 70 cutting-edge papers on human-related analysis and vision-language understating. She was the recipient of Best Paper Award of ACM MM 2021 and 2013, Best Demo Award of ACM MM 2012. She was the Champion of CVPR 2017 Look Into Person Challenge and the organizer of the ECCV 2018, ICCV 2019, CVPR 2021, CVPR 2022 and ACM MM 2022 Person in Context Challenges. She is the Associate Editor of IEEE TMM and TCSVT.

Yulu Gao is a PhD student from Computer Science and Engineering, Beihang University. He received the Bachelor degree from Beihang University. His research interests include object detection and visual tracking.

Aixi Zhang received the Mphil degree from the Hong Kong University of Science and Technology, Hong Kong, China, in 2014. He is now a senior researcher at Taobao (China) Software Co. Ltd., Alibaba Group. His research interests include computer vision, HOI detection and multimodal video understanding. He has published 4 papers at top conferences and journals including NIPS, CVPR, ACM MM and TIP. He was the Champion of CVPR 2021 ActivityNet Homage Challenge and CVPR 2020 DeepFashion2 Fashion Retrieval Challenge.

Zhimin Li received the B.S. degree from Wuhan University of Technology, Wuhan, China, in 2019, and the M.S. degree from Huazhong University of Science and Technology, Wuhan, China, in 2022. His research interests include video analysis and scene understanding. To date, he has published in these fields with some peer-reviewed technical papers at premium conferences including AAAI, CVPR, ACM MM, etc. He currently serves as the reviewer of several conferences, including CVPR, ECCV, etc..

Fei Wang is the Director of SenseTime Intelligent Automotive Group. He is the head of SenseAuto-Parking engineering and SenseAuto-Cabin research. He has published 30+ papers at CVPR/NIPS/ICCV and gained over 4000 Google Scholar Citations during the last few years. Fei obtained his Bachelor's degree and Master's degree from Beijing University of Posts and Telecommunications. Currently, he is a Ph.D. student at the University of Science and Technology of China. His research interests include Automotive Drive System, AI Chip, Deep Learning, etc.

Bo Li is currently a Changjiang Distinguished Professor of School of Computer Science and Engineering, Beihang University. He is a recipient of The National Science Fund for Distinguished Young Scholars. He is currently the dean of AI Research Institute, Beihang University. He is the chief scientist of National 973 Program and the principal investigator of the National Key Research and Development Program. He has published over 100 papers in top journals and conferences and held over 50 domestic and foreign patents.