DIRV: Dense Interaction Region Voting for End-to-End Human-Object Interaction Detection

Hao-Shu Fang[1*], Yichen Xie[1*], Dian Shao[2], Cewu Lu[1†]
[1]Shanghai Jiao Tong University [2]The Chinese University of Hong Kong
fhaoshu@gmail.com, xieyichen@sjtu.edu.cn, sd017@ie.cuhk.edu.hk, lucewu@sjtu.edu.cn

Abstract

*Recent years, human-object interaction (HOI) detection has achieved impressive advances. However, conventional two-stage methods are usually slow in inference. On the other hand, existing one-stage methods mainly focus on the union regions of interactions, which introduce unnecessary visual information as disturbances to HOI detection. To tackle the problems above, we propose a novel one-stage HOI detection approach **DIRV** in this paper, based on a new concept called interaction region for the HOI problem. Unlike previous methods, our approach concentrates on the densely sampled interaction regions across different scales for each human-object pair, so as to capture the subtle visual features that is most essential to the interaction. Moreover, in order to compensate for the detection flaws of a single interaction region, we introduce a novel voting strategy that makes full use of those overlapped interaction regions in place of conventional Non-Maximal Suppression (NMS). Extensive experiments on two popular benchmarks: V-COCO and HICO-DET show that our approach outperforms existing state-of-the-arts by a large margin with the highest inference speed and lightest network architecture. Our code will be made publicly available.*

1. Introduction

Human-object interaction (HOI) detection aims to recognize and localize the interactions between human-object pairs (*e.g.* sitting on a chair, riding a horse, eating an apple, *etc.*). As a fundamental task of image semantic understanding, it plays a vital role in many other computer vision fields such as image captioning [9, 22], visual question answering [14, 25] and action understanding [34, 28].

For HOI detection, almost all previous methods emphasized the importance of the *union regions* of an interaction, which covers the whole human, object and intermediate context. For instance, existing two-stage algorithms commonly crop the union region of a human-object pair and then embed its visual features [11, 7, 17], while recent one-stage methods aim to regress this union region with keypoints [18, 31] or anchor boxes [12] and use it to associate the target human and object.

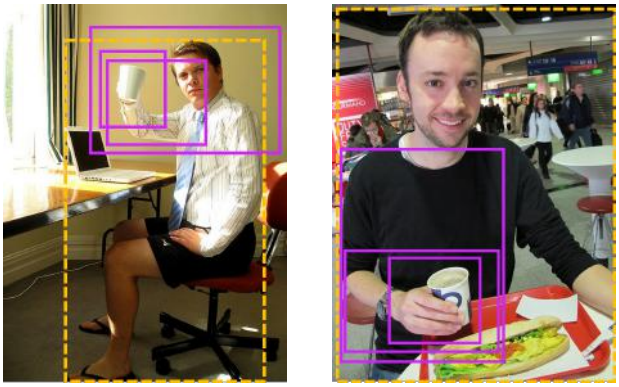


Figure 1. **Union Regions *vs* Interaction Regions:** Conventional approaches usually pays attention to the *union region* (*dashed yellow*), which contains too much redundant information. Instead, we propose a method focusing on *interaction regions* (*solid violet*) with different scales. In above two figures, despite distinct human/object poses, *interaction regions* cover the most critical segments containing the cups, hands or arms, when detecting *holding a cup*.

However, we find that such emphasis on union region is *counter-intuitive* for human beings. In practice, it is not necessary to observe the whole union region before making decisions in most situations. For instance, when asked to determine whether a man is holding a cup, we only need to notice his hands but never care about where his feet are. That's to say, humans can easily target the human-object pair of an HOI, without the needs of being told the union regions. Based on these observations, we propose a new recognition unit for HOI detection, called *interaction region*. The interaction region denotes the region that covers

* Equal contribution. Names in alphabetical order.
† Cewu Lu is the corresponding author.

the minimal area of human and object crucial for recognizing the interaction. An example is given in Fig. 1. In this case, an upper-body region that contains a cup and hand would be more distinguishable than the union region.

To this end, we propose a novel one-stage HOI detector that concentrates on the interaction regions of human-object interactions. We assume that these regions are highly informative to determine the interaction category and human-object relative spatial configuration. To fully utilize the interaction regions for HOI detection, three main technical challenges identified as follows need to be addressed beforehand.

Challenge 1: how do we decide the interaction regions? Although recent work provided part-level action labels [16], we tend to seek a more general and simpler HOI detector without the need for extra annotations. Empirically, we consider that those human parts closer to the object are more likely to take an indispensable effect on the interaction, and so are the object parts. For simplicity, we consider some rectangle regions, which cover both some parts of the human and object, as interaction regions. A natural idea comes by applying the dense anchor boxes in one-stage object detection models to represent these regions. To achieve that, we set three overlapping thresholds between anchor boxes and human bounding boxes, object bounding boxes as well as union regions. We apply a dense interaction region selection manner, where all anchors satisfying these three thresholds are regarded as interaction regions.

Challenge 2: an anchor box may be regarded as the interaction region for multiple different HOIs. Unlike object detection, this situation appears frequently in HOI detection. Under this condition, the anchor box needs to predict multiple HOI labels and corresponding object locations, where the number is unfixed. This poses extra challenges for network design and final result association. Therefore, we match each anchor box with only one unique interaction. In addition, there inevitably exists some missed positive interactions within the popular datasets. We develop a novel *ignorance loss* based on classical focal loss [20] to address these problems.

Challenge 3: single interaction region may lead to ambiguity or misrepresentation. HOI recognition relies on very subtle visual cues in interaction regions. Some visual features are even ambiguous, leading to the fragile result from a single anchor. For this reason, we propose a novel *voting strategy*. Each anchor only contributes a little to the final location and classification prediction. For each interaction type, a *probability distribution* is established for the relative location between each human-object pair by fusing the prediction results of different anchors. This *dense anchor voting strategy* can remarkably elevate the fault-tolerance of each anchor and achieve a robust final prediction.

Extensive experiments show that our one-stage approach, **DIRV** (**D**ense **I**nteraction **R**egion **V**oting), outperforms existing state-of-the-art models on two popular benchmarks, achieving both *higher accuracy* and *faster speed*.

2. Related Work

Human-object interaction (HOI) detection is formally defined as retrieving $\langle$*human, verb, object*$\rangle$ triplets from images. Previous methods mainly employed a two-stage strategy. In the first stage, a pre-trained object detector [19, 27] localized both humans and objects within the image. In the second stage, a classification network recognized the interaction categories for each human-object pair. Most work focused on the improvement of the second stage. Some early work [10] simply extracted features from each human or object instance. This method suffered from lack of contextual information. Afterwards, more information was taken into account rather than instance appearance, including spatial location [2, 7, 26], human pose [5, 17, 33], word embedding [1, 24], segmentations [35, 4] and human part label [16]. Yet, these two-stage methods typically need to detect all human-object pairs, making their inference time grow quadratically with instance number. Furthermore, these approaches usually adopted a heavy network for classification, which led to considerable computation overhead.

To tackle these drawbacks, some recent work developed one-stage HOI detectors. Liao *et.al.* [18] and Wang *et.al.* [31] posed HOI detection as a keypoint detection and grouping problem. Despite their impressive efficiency and accuracy, the interaction keypoints had no apparent characteristics in visual patterns so the networks were not easy to train. Kim *et.al.* [12] designed an anchor-based one-stage algorithm to regress the union region of human and object. However, as aforementioned, union region prediction is not straight-forward and single anchor's prediction is fragile.

Unlike all the above methods, our method makes full use of visual patterns within interaction regions across different scales, allowing a promising accuracy without the help of any other proposals or annotations. The one-stage strategy and concise network architecture also bring greatly improved running time and space efficiency.

3. Methods

In this section, we introduce our proposed **DIRV** (Dense Interaction Regions Voting) framework for human-object interaction (HOI) detection. The problem formulation is firstly explained in Sec. 3.1. Then, we present the network architecture of our detector in Sec. 3.2. Afterwards, the inference protocol based on *voting strategy* is shown in Sec. 3.3. Finally, we demonstrate how to train our deep neural network model in Sec. 3.4.

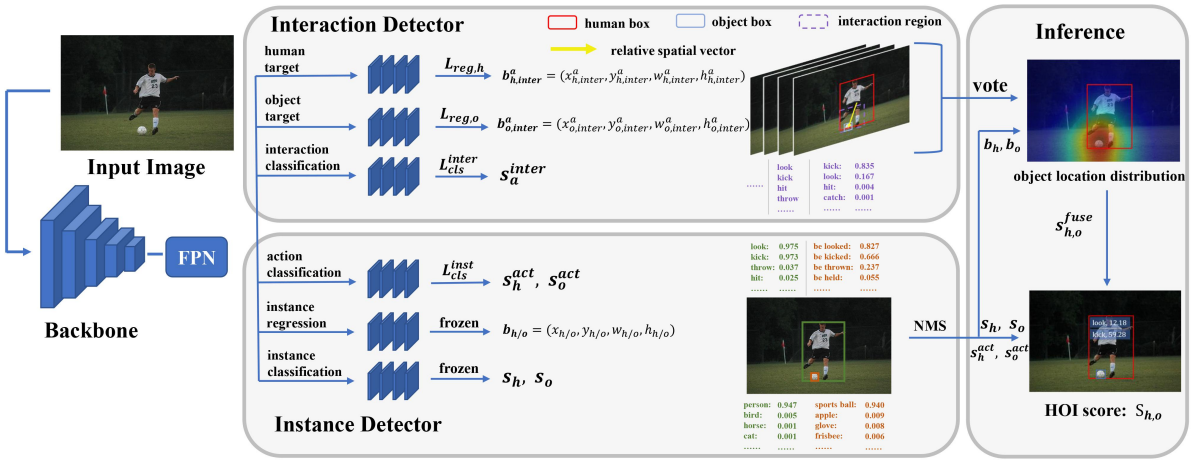


Figure 2. **Overview of our DIRV Framework:** It is composed of two components: *Interaction Detector* and *Instance Detector*. For each interaction region, a relative spatial vector is obtained by regressing the human and object bounding boxes. During inference, results of interaction regions vote for an *object location distribution*, from which HOI score is derived.

3.1. Formulation

Typically, HOI detection aims to fetch a $\langle b_h, v, b_o \rangle$ triplet for each interaction within a single image x, where b_h, b_o denote the bounding box of human h and object o separately, while v denotes the human action. Without considering external input like human poses [6], conventional two-stage HOI detectors formulate the problem as

$$\begin{aligned} &\mathcal{H}, \mathcal{O} = d(\mathbf{f_x}), \\ &v_i = g(b_h, b_o, \mathbf{f_x}), \forall h \in \mathcal{H}, \forall o \in \mathcal{O}, \end{aligned} \tag{1}$$

where $d(\cdot)$ is a vanilla object detector, $g(\cdot)$ is the verb classifier for a human-object pair, $\mathbf{f_x}$ is the appearance feature of the whole image x and $\mathcal{H}$,$\mathcal{O}$ are detected humans and objects. Since the input of $g(\cdot)$ relies on the output of $d(\cdot)$, these two processes cannot run in parallel and $g(\cdot)$ would face the combinatorial explosion problem. On the contrary, we reformulate HOI detection as

$$\begin{aligned} &\mathcal{H}, \mathcal{O} = d(\mathbf{f_x}), \\ &\langle T(b_h), v, T(b_o) \rangle = g(\mathbf{f_x}), h \in \mathcal{H}, o \in \mathcal{O}, \end{aligned} \tag{2}$$

where $T(\cdot)$ is a target indicator that links the verb to a detected human-object pair. By doing so, we can run these two processes simultaneously.

Further, we do not adopt the common practice of Non-Maximum Suppression (NMS) when retrieving the $\langle T(b_h), v, T(b_o) \rangle$. In contrast, we propose a different strategy, *voting*, to handle the prediction of different *interaction regions*. Predictions based on every anchor's visual features are fully utilized instead of being suppressed. The final HOI prediction comes from the combination of each interaction region through voting. To sum up, our algorithm is formulated as Eq. 3:

$$\begin{aligned} &\mathcal{H}, \mathcal{O} = d(\mathbf{f_x}), \\ &\langle T(b_h^i), v^i, T(b_o^i) \rangle = g(\mathbf{f_x^{a_i}}), i \in \{1, 2, \ldots, N\}, \\ &\langle T(b_h), v, T(b_o) \rangle = vote(\{\langle T(b_h^i), v^i, T(b_o^i) \rangle\}_{i \in \{1,\ldots,N\}}), \end{aligned} \tag{3}$$

where $\langle T(b_h^i), v^i, T(b_o^i) \rangle$ is the prediction based on anchor a_i. N is the number of interaction regions for this interaction. We show how we obtain $\mathcal{H}, \mathcal{O}$ and $\langle T(b_h^i), v^i, T(b_o^i) \rangle$ for each anchor in Sec. 3.2. $vote(\cdot)$ is the voting strategy, which is elaborated in Sec. 3.3.

3.2. Dense Interaction Region Detector

Our network structure is illustrated in Fig. 2. The model is composed of two components: an instance detector and an interaction detector. Each of them contains three parallel sub-branches, which share the feature map of the Feature Pyramid Network. We first explain the instance detector for $\mathcal{H}, \mathcal{O}$ and then the interaction detector for $\langle T(b_h^i), v^i, T(b_o^i) \rangle$.

3.2.1 Instance Detector

The instance detector mainly helps instance localization and supports the detection of none object actions, *e.g. walking*. It contains three sub-branches: *instance classification branch*, *instance regression branch* and *instance action classification branch*.

The instance regression and classification branches follow the standard setting in most object detection networks, which regress instance bounding boxes based on anchors as well as classify these instances. Interactions are not considered in these two branches.

Beyond these two branches, an instance action classification branch plays an auxiliary role in interaction classifi-

cation. It predicts the action scores of humans and objects, helping the association of human-verb-object pair. The actions of humans and objects are treated separately, *e.g.*, *hold* and *be held* are classified as two different actions. If there are C_h human actions and C_o object actions, the classification gives two scores $s_h^{act} \in \mathbb{R}^{C_h}$ and $s_o^{act} \in \mathbb{R}^{C_o}$. The anchor settings follow standard object detection and only those positive anchors involved in at least one interaction are taken into account when calculating loss.

3.2.2 Interaction Detector

The interaction detector serves as the key of our proposed architecture, **DIRV**. It directly predicts the interaction v^i and the target $\langle T(b_h^i), T(b_o^i)\rangle$ that indicates the corresponding human-object pair from the subtle visual features in *interaction regions*. We first clarify our *methodology*, followed by two key learning techniques: *interaction region decision* and *ignorance loss*.

Methodology: To retrieve the $\langle T(b_h^i), v^i, T(b_o^i)\rangle$ triplet, we design three parallel sub-branches: *interaction classification branch*, *human target branch* and *object target branch* for predicting $v^i, T(b_h^i)$, and $T(b_o^i)$ separately.

The interaction classification branch classifies the interaction type v^i within the interaction region (*i.e.* the anchor). It obtains an interaction score prediction $s_{a_i}^{inter} \in \mathbb{R}^C$ for each interaction region a_i, where C is the number of interaction categories.

For human and object targets $T(b_h^i)$ and $T(b_o^i)$, it is difficult to directly link the verb to the detected human and object given by the *instance detector* since the detection branch run in parallel. Thus, we propose an intuitive yet effective solution. The human target branch regresses the human bounding box $b_{h,inter}^{a_i} = (x_{h,inter}^{a_i}, y_{h,inter}^{a_i}, w_{h,inter}^{a_i}, h_{h,inter}^{a_i})$ from the anchor $b_a^{a_i} = (x_a^{a_i}, y_a^{a_i}, w_a^{a_i}, h_a^{a_i})$, where $(x_{h,inter}^{a_i}, y_{h,inter}^{a_i})$ is its bounding box center. Similarly, the object target branch regresses the object bounding box $b_{o,inter}^{a_i} = (x_{o,inter}^{a_i}, y_{o,inter}^{a_i}, w_{o,inter}^{a_i}, h_{o,inter}^{a_i})$. These predicted human and object bounding boxes serve as the target indicators $T(b_h^i)$ and $T(b_o^i)$. We can easily link the verb v^i to the detected human and object box b_h^i, b_o^i during inference via simple post processing (*e.g.,* IoU matching), which is introduced in Sec.3.3.

Interaction Region Decision: As explained before, the interaction regions should cover both parts of interacting human and object. With different scales, these regions may provide important visual features of different levels. Interestingly, we find that such a setting naturally matches the characteristic of anchor boxes $\mathcal{A}$. An anchor box $a_j \in \mathcal{A}$ serves as an interaction region of interaction I_i so long as it satisfies the following overlapping requirement:

$$O_i^j = \mathbb{1}\left(IoU(a_j, \hat{b}_u^i) > t_u\right) \cdot \mathbb{1}\left(\frac{a_j \cap \hat{b}_h^i}{\hat{b}_h^i} > t_h\right) \cdot \mathbb{1}\left(\frac{a_j \cap \hat{b}_o^i}{\hat{b}_o^i} > t_o\right) \tag{4}$$

where $\hat{b}_h^i, \hat{b}_o^i$ are the ground-truth human/object bounding box of a possible interaction pair I_i. $\hat{b}_u^i$ is the union region box of interaction I_i, which is the smallest box that completely covers both $\hat{b}_h^i, \hat{b}_o^i$. t_u, t_h, t_o are three thresholds. We set them as $t_u = t_h = t_o = 0.25$, which is analyzed in ablation study.

With the requirement above, single anchor box may serve as the interaction region of multiple interactions, which impedes the human/object regression. Thus, we define a *overlapping level* metric to ensure that an anchor box corresponds to at most a unique interaction, *i.e.,*

$$\hat{O}_i^j = IoU(a_j, \hat{b}_u^i) + \sqrt{\frac{a_j \cap \hat{b}_h^i}{\hat{b}_h^i} \cdot \frac{a_j \cap \hat{b}_o^i}{\hat{b}_o^i}}. \tag{5}$$

If multiple interactions are matched with the same anchor box, it will associate with interaction I_k where $\hat{O}_k^j = \max_i \left\{\hat{O}_i^j | O_i^j = 1\right\}$ so each anchor has at most one ground-truth in regression.

Ignorance Loss: For human/object target branch, we just follow many anchor-based object detection methods to apply the standard smooth L_1 loss between predicted $b_{h,inter}^{a_i}/b_{o,inter}^{a_i}$ and ground-truth $\hat{b}_h^i/\hat{b}_o^i$ on their loss functions $\mathcal{L}_{reg,h}/\mathcal{L}_{reg,o}$ for interaction region a_i.

Yet, standard focal loss is not applicable for interaction classification branch because of the following two reasons: Firstly, the receptive field of an anchor may contain multiple different interactions. Secondly, HOI detection datasets have much more missed positive samples than object detection datasets. These cause serious confusion during training.

We propose a novel *ignorance loss* based on vanilla focal loss [20] to address both difficulties above. We eliminate the influence of missed unlabelled interactions by removing the background loss *i.e.* anchors associated with none interactions *don't* take effect in learning.

Further, as a solution to the multiple interactions problem, we modify the ground-truth targets of foreground anchors as below. For anchor a_j, if there exist multiple interactions $\{I_i\}$ within current anchor where $O_i^j = 1$, we set the target label as

$$t_j^c = \begin{cases} 1 & I_k^c = 1, \hat{O}_k^j = \max_i\{\hat{O}_i^j | O_i^j = 1\} \\ 0 & I_i^c = 0, \forall i, O_i^j = 1 \\ ignored & others \end{cases} \tag{6}$$

where t_j^c is the target label of interaction category c for anchor a_j. $I_i^c = 1$ denotes interaction I_i is positive for category c, else $I_i^c = 0$. The above equation means that we ignore the classification loss for those interaction categories exist but not dominant in an anchor.

3.3. Voting Based Model Inference

Our model makes inference by combining the prediction results of different interaction regions. Each interaction region contributes to the final interaction recognition with the *weighted localization score* as weight. The inference process is divided into three steps as follows.

3.3.1 Parallel Inference

All six sub-branches work in parallel during inference, which dramatically reduces the inference time. From *instance detector*, a set of human $\mathcal{H}$ and object $\mathcal{O}$ ($\mathcal{H} \subset \mathcal{O}$) candidates are generated after NMS. For each human instance, we get its bounding box $b_h \in \mathbb{R}^4$, instance classification score $s_h \in \mathbb{R}$ and instance action classification score $s_h^{act} \in \mathbb{R}^{C_h}$. $s_h \in \mathbb{R}$ is a scalar since an instance can only be classified as a unique object category with highest score (here is *human*) while $s_h^{act} \in \mathbb{R}^{C_h}$ is a C_h-d vector. Similarly, we obtain bounding box $b_o \in \mathbb{R}^4$, instance classification score $s_o \in \mathbb{R}$ and instance action classification score $s_o^{act} \in \mathbb{R}^{C_o}$ for each object.

In *interaction detector*, it fetches a triplet of $(b_{h,inter}^{a_i}, s_{a_i}^{inter}, b_{o,inter}^{a_i})$ from each interaction region a_i, where $b_{h,inter}^{a_i}, b_{o,inter}^{a_i} \in \mathbb{R}^4$ are the human/object target bounding boxes and $s_{a_i}^{inter} \in \mathbb{R}^C$ is the interaction classification score for each interaction region. Here, we should have $C = C_h = C_o$ after eliminating interactions with none objects.

3.3.2 Object Location Estimation

We retrieve the $\langle b_h, v, b_o \rangle$ triplet in a human-centric manner. For each interaction region a_j, we first try to match it with a human instance $h^{a_j} \in \mathcal{H}$ based on the overlapping metric, that is

$$\begin{aligned} &IoU(a_j, b_h^{a_j}) = \max_h IoU(a_j, b_h), \\ &h \in \mathcal{H}, \frac{a_j \cap b_h}{b_h} > t_h \end{aligned} \tag{7}$$

where b_h is the human bounding box and t_h is the threshold same as that in Eq. 4. If no human instance meets the requirement, this interaction region is abandoned.

After matching the interaction region to a detected human instance, we then search its corresponding object instance. A natural thinking is to match the object like Eq.7. However, we found that the location of object is usually not accurate enough. To improve the robustness, we build a probability distribution for the object location based on the prediction result. Referring to [8], we model it with a 2-d Gaussian distribution:

(a) eat (b) talk on phone

(c) surf (d) throw

Figure 3. **Object Location Distribution:** we visualize the target object location distribution for some human instances of several categories. Our voting strategy accurately localizes the objects in these interactions.

$$p_{a_j}(x_o, y_o) = e^{-\frac{\left|\left|v_{o|h}^{a_j} - \mu_{o|h}^{a_j}\right|\right|^2}{2 \cdot \sigma^2}} \tag{8}$$

where $v_{o|h}^{a_j}$ and $\mu_{o|h}^{a_j}$ are the relative object locations scaled by anchor width and height:

$$\begin{aligned} v_{o|h}^{a_j} &= \left(\frac{x_o - x_h^{a_j}}{w_a^{a_j}}, \frac{y_o - y_h^{a_j}}{h_a^{a_j}} \right), \\ \mu_{o|h}^{a_j} &= \left(\frac{x_{o,inter}^{a_j} - x_{h,inter}^{a_j}}{w_a^{a_j}}, \frac{y_{o,inter}^{a_j} - y_{h,inter}^{a_j}}{h_a^{a_j}} \right), \end{aligned} \tag{9}$$

and the standard deviation σ is a hyper-parameter, which is set as 0.9 in our experiments. Its influence is analyzed in supplementary material.

After obtaining the object location distribution, we weight it by *interaction classification score* $s_{a_j}^{inter}$ as below.

$$s_{a_j}^{loc}(x_o, y_o) = s_{a_j}^{inter} \cdot p_{a_j}(x_o, y_o) \tag{10}$$

where (x_o, y_o) is the center of object bounding box. Until now, we obtain the *weighted localization scores* $s_{a_j}^{loc}(x, y) \in \mathbb{R}^C$ for all C interaction categories.

3.3.3 Voting Based Region Fusion

By fusing *weighted localization scores* of interaction regions associated with same human instance b_h, a *human-*

centric object location distribution s_h^{fuse} is computed with our voting strategy:

$$s_h^{fuse}(x,y) = \sum_{a_j \in \mathcal{A}_h} s_{a_j}^{loc}(x,y), \tag{11}$$

where $\mathcal{A}_h = \{a_j\}_{h^{a_j}=h}$ is set of interaction regions associated with human instance h. We visualize some examples of the fused distribution in Fig. 3.

Eventually, we are now able to score a human-object pair using this distribution. For each interaction region, we first associate it with a detected object instance o^{a_j}, like Eq. 7.

$$\begin{aligned} &p_{a_j}(x_o^{a_j}, y_o^{a_j}) = \max_o p_{a_j}(x_o, y_o), \\ &o^{a_j} \in \mathcal{O}, \frac{a_j \cap b_o}{b_o} > t_o. \end{aligned} \tag{12}$$

Then, Eq. 11 is rewritten for each specific human-object pair.

$$s_{h,o}^{fuse} = \sum_{a_j \in \mathcal{A}_{h,o}} s_{a_j}^{loc}(x_o, y_o) \tag{13}$$

where $\mathcal{A}_{h,o}$ denotes all the interaction regions $\{a_j\}$ associated human-object pair (b_h, b_o) where $(b_h, b_o) = (b_h^{a_j}, b_o^{a_j})$. Thus, the final HOI score for a human-object pair (b_h, b_o) can be derived as

$$S_{h,o} = s_h \cdot s_o \cdot (s_h^{act} + s_o^{act}) \cdot s_{h,o}^{fuse} \tag{14}$$

where $s_h, s_o, s_h^{act}, s_o^{act}$ have been explained in section *Parallel Inference*. When no object is involved, we simply define $S_h = s_h \cdot s_h^{act}$. The HOI scores are not normalized because we only care about their relative value for the same interaction category.

The time complexity of voting is $O(|\mathcal{A}_{pos}|)$, where $\mathcal{A}_{pos} = \bigcup_{h,o} \mathcal{A}_{h,o}$ is the set consisting of all interaction regions associated with any interactive human-object pairs. The size is not very large and it is easy to compute in parallel, so only a little CPU overhead is introduced.

3.4. Model Training

During training, the backbone, feature pyramid network and instance classification/regression branches are frozen with COCO pre-trained weight [30]. The final loss is the sum of loss functions for other four sub-branches in Fig. 2.

$$\mathcal{L} = \mathcal{L}_{reg,h} + \mathcal{L}_{reg,o} + \mathcal{L}_{cls}^{inter} + \mathcal{L}_{cls}^{inst} \tag{15}$$

In *interaction detector*, $\mathcal{L}_{reg,h}, \mathcal{L}_{reg,o}$ are the smooth L_1 losses for *human and object target branches* separately. $\mathcal{L}_{cls}^{inter}$ is our *ignorance loss* for *interaction classification branch*. We follow focal loss [20] to set $\alpha = 0.25, \gamma = 2.0$. In *instance detector*, $\mathcal{L}_{cls}^{inst}$ is standard binary cross-entropy loss for *instance action classification branch*.

4. Experiments

In this section, we carry out comprehensive experiments to demonstrate the superiority of our proposed **DIRV**. Firstly, we introduce two benchmarks in Sec. 4.1 and model implementation details in Sec. 4.2. Then, we compare the performance of our model with other state-of-the-art approaches in Sec. 4.3. Finally, effect of some crucial configurations are examined with ablation study in Sec. 4.4

4.1. Dataset and Metric

Dataset We evaluate our method on two popular datasets: **V-COCO** [10] and **HICO-DET** [3]. V-COCO dataset is a subset of COCO [21] with extra interaction labels. It contains 10,346 images (2,533 for training, 2867 for validation and 4,946 for testing). Each person in these images is annotated with 29 action categories, 4 of which (*stand, smile, walk, run*) have no object. HICO-DET is a large dataset for HOI detection by augmenting HICO dataset [3] with instance bounding box annotations. This dataset includes 38,118 images for training and 9,658 images for testing. It is labelled with 600 HOI types over 117 verbs and 80 object categories.

Metric We adopt the popular evaluation metric for HOI detection: *mean average precision (mAP)*. A prediction is true positive only when the HOI classification result is accurate as well as bounding boxes of human and object both have IoUs larger than 0.5 with reference to ground-truth. Specifically, we follow prior works to report *Scenario 1 role mAP* on V-COCO dataset.

4.2. Implementation Details

For HOI detection, we use EfficientDet-d3 [30] as the backbone due to its effectiveness and efficiency. The backbone is pre-trained on COCO dataset. The *instance classification and regression branches* are also initialized with the COCO pre-trained weight, which is frozen during training. We apply random flip and random crop data augmentation approaches to our model. Adam optimizer [13] is employed to optimize the loss function. We set the learning rate as 1e-4 with a batch size of 32. All experiments are carried out on NVIDIA RTX2080Ti GPUs.

4.3. Results and Comparison

We compare our proposed **DIRV** with other state-of-the-art methods on V-COCO (Tab. 1) and HICO-DET (Tab. 2) datasets. It is noticeable that many state-of-the-art models utilize other additional features like human poses and language priors. These methods require additional data, annotations or models, which are quite exhaustive to collect. For fairness, we *do not* take them (gray ones in both Tab. 1,2) into account in our comparison. What's more, unlike many

Table 1. **Results on V-COCO:***Proposal* shows whether it needs object detection beforehand. For *Additional*, *P,B,L* denotes human pose, human body part states and language priors respectively, which are utilized in prior methods.

Method	**Proposal**	**Additional**	**mAP**$_{role}$	**Inference Time** (ms)	**Params** (M)
RP_DC_D [17]	✓	P	47.8	513	64M
PMFNet [33]	✓	P	52.0	253	179M
ConsNet [23]	✓	P+L	53.2	-	-
MLCNet [29]	✓	P+B+L	55.2	-	-
InteractNet [8]	✓	✗	40.0	145	$35M$
iCAN [7]	✓	✗	44.7	204	$89M$
Zhou *et.al.* [35]	✓	✗	48.9	-	$620M$
VSGNet [32]	✓	✗	51.8	312	$59M$
UnionDet [12]	✗	✗	47.5	78	$35M$
IP-Net [31]	✗	✗	51.0	-	$195M$
DIRV (ours)	✗	✗	**56.1**	**68**	**12M**

Table 2. **Results on HICO-DET:** *Proposal* shows whether it needs object detection beforehand. For *Additional*, *P,B,L* denotes human pose, human body part states and language priors respectively, which are utilized in prior methods.

Method	**Proposal**	**Additional**	**Default**			**Known Object**		
			Full	Rare	Non-Rare	Full	Rare	Non-Rare
RP_DC_D [17]	✓	P	17.03	13.42	18.11	19.17	15.51	20.26
PMFNet [33]	✓	P	17.46	15.65	18.00	20.34	17.47	21.20
MLCNet [29]	✓	P+B+L	17.95	16.62	18.35	22.28	20.73	22.74
Functional [1]	✓	L	21.96	16.43	23.62	-	-	-
ConsNet [23]	✓	P+L	22.15	17.12	23.65	-	-	-
InteractNet [8]	✓	✗	9.94	7.16	10.77	-	-	-
iCAN [7]	✓	✗	14.84	10.45	16.15	16.26	11.33	17.73
UnionDet [12]	✗	✗	17.58	11.72	19.33	19.76	14.68	21.27
IP-Net [31]	✗	✗	19.56	12.79	21.58	22.05	15.77	23.92
PPDM-DLA [18]	✗	✗	20.29	13.06	22.45	23.09	16.14	25.17
DIRV (S1)	✗	✗	**21.40**	**15.52**	**23.15**	**24.53**	**18.66**	**26.28**
DIRV (S2)	✗	✗	**21.78**	**16.38**	**23.39**	**25.52**	**20.84**	**26.92**

existing two-stage approaches, our method does not rely on object proposals, which significantly elevates its compatibility.

For V-COCO dataset (Tab. 1), we follow prior works to ignore the class *point* since it has too few samples. Compared to prior arts, our approach outperforms them in accuracy significantly. It also has a fastest inference speed and a least parameter number.

For HICO-DET dataset, we report the results on two different settings: Default and Known Objects. Meanwhile, we explore two possible interaction classification strategies:

S1 The interaction classification branch directly recognizes different verb-object pairs *e.g. eating apples*, as in [7, 17].

S2 Only verb categories *e.g. eating* are classified in interaction classification branch, which are associated with object categories *e.g. apple* based on the results of instance classification branch, as in [12].

As shown in Tab. 2, our approach exceeds all existing methods in accuracy notably. Since Hourglass-104 backbone has more than $10\times$ parameters than ours, we compare with DLA-34 based PPDM for fairness despite it still has $2\times$ more parameters. We can see that S2 brings a more promising performance. The reason may be that it reduces the number of categories in interaction classification branch, which elevates the accuracy. What's more, it saves the space overhead, allowing a larger batch size during training and improving the training stability. Our approach has a superiority in time and space complexity. Due to limited space, we do not list the model parameters and inference time in the table.

Two prior arts share some common insights with us. InteractNet [8] localizes objects based on single human appearance. UnionDet [12] is another anchor-based one-stage HOI detection approach, focusing on *union regions*. How-

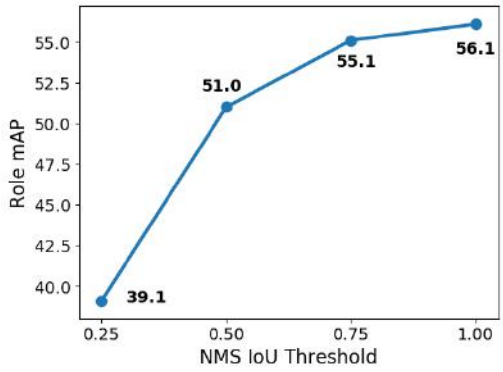


Figure 4. **Ablation Study for Voting Strategy:** The mAP_{role} decreases as the IoU threshold for NMS grows. There is actually no NMS when IoU threshold is 1.

ever, we surpass their performance by a large margin on both datasets, which proves the effectiveness of our concentration on interaction regions and our dense interaction region voting strategy.

In the supplementary materials, we show some qualitative results of our network, which is further analyzed with visualization.

4.4. Ablation Study

In this section, we dig into the influence of different modules in our **DIRV**. For simplicity, all results here are for V-COCO dataset. Analysis of more components are available in the supplementary materials.

Table 3. **Interaction Region Overlapping Thresholds:** t_u, t_h, t_o denote the thresholds in Eq. 4. The interaction regions become denser as these three thresholds decrease.

$\mathbf{t_h}$	$\mathbf{t_o}$	$\mathbf{t_u}$	$\mathbf{mAP}_{role}$
0.5	0.5	0.5	55.0
0.25	0.25	0.5	55.2
0.25	0.25	0.25	**56.1**

Interaction Regions Overlapping Thresholds We set interaction regions in a dense manner for human-object pairs. The overlapping thresholds in Eq. 4 is examined in this part. Results in Tab. 3 certificate this dense manner, which can make full use of the visual features.

Voting Strategy We examine the superiority of our voting strategy by adding a NMS module for interaction regions, which weakens the effect of voting. In Fig. 4, we set different IoU thresholds for NMS and the performance drops as the value of those thresholds decreases (when IoU threshold is 1, NMS takes no effect). It reveals that interaction regions of different scales all contribute to the final detection though some of their classification scores may not be very high.

Table 4. **Loss Function for Interaction Classification**

Loss Function	$\mathbf{mAP}_{role}$
Focal Loss [20]	54.8
Foreground Loss [12]	54.0
Ignore Loss (ours)	**56.1**

Ignorance Loss We look into the effect of loss function in *interaction classification branch*. We test the performance with vanilla focal loss, foreground loss in [12] and our proposed *ignorance loss*. Results in Tab. 4 verify our superiority since it can help dealing with region overlapping and missed positive labels.

Backbone We apply a novel backbone [30] to our model, which has never been utilized for HOI detection.

We separately carry out experiments with EfficientDet-d1, d2, d3 and d4. To our surprise, we find that the heavier backbone doesn't certainly lead to better HOI detection performance, according to the results in Tab. 5.

We also reproduce another anchor-based one-stage algorithm UnionDet [12] with EfficientDet-d3 backbone. Results in Tab. 5 reveals that our DIRV surpasses it because of our novel design in methodology, instead of the backbone improvement.

Table 5. **Ablation Study for Backbones:** We compare the performance of our DIRV and another anchor-based method UnionDet [12] with different backbones.

Method	**Backbone (PARAMs)**	$\mathbf{mAP}_{role}$
UnionDet	ResNe50-FPN (34M)	47.5
UnionDet	EfficientDet-d3 (12M)	49.2
DIRV (ours)	EfficientDet-d1 (6.6M)	46.8
DIRV (ours)	EfficientDet-d2 (8.1M)	49.4
DIRV (ours)	EfficientDet-d3 (12M)	**56.1**
DIRV (ours)	EfficientDet-d4 (21M)	54.3

5. Conclusion

In this paper, we present a novel one-stage HOI detection framework. It detects HOI in an intuitive manner by concentrating on the *interaction regions*. To compensate for the detection flaws of single interaction region, a *voting strategy* is applied as an alternative to conventional NMS. Our method outperforms all existing approaches without any additional features or proposals. Due to the one-stage structure and simple network architecture, our method reaches a very high efficiency with least model parameters compared to other state-of-the-art approaches. In the future, we will try to incoporate the part-level knowledge [15] into our framework.

References

[1] A. Bansal, S. S. Rambhatla, A. Shrivastava, and R. Chellappa. Detecting human-object interactions via functional generalization. *CoRR*, abs/1904.03181, 2019. 2, 7, 10

[2] Y.-W. Chao, Y. Liu, X. Liu, H. Zeng, and J. Deng. Learning to detect human-object interactions. In *2018 ieee winter conference on applications of computer vision (wacv)*, pages 381–389. IEEE, 2018. 2

[3] Y.-W. Chao, Z. Wang, Y. He, J. Wang, and J. Deng. HICO: A benchmark for recognizing human-object interactions in images. In *Proceedings of the IEEE International Conference on Computer Vision*, 2015. 6

[4] H. Fang, G. Lu, X. Fang, J. Xie, Y. Tai, and C. Lu. Weakly and semi supervised human body part parsing via pose-guided knowledge transfer. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 70–78, 2018. 2

[5] H.-S. Fang, J. Cao, Y.-W. Tai, and C. Lu. Pairwise body-part attention for recognizing human-object interactions. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 51–67, 2018. 2

[6] H.-S. Fang, S. Xie, Y.-W. Tai, and C. Lu. Rmpe: Regional multi-person pose estimation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2334–2343, 2017. 3

[7] C. Gao, Y. Zou, and J.-B. Huang. ican: Instance-centric attention network for human-object interaction detection. *arXiv preprint arXiv:1808.10437*, 2018. 1, 2, 7, 10

[8] G. Gkioxari, R. Girshick, P. Dollár, and K. He. Detecting and recognizing human-object interactions. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 5, 7, 10

[9] L. Guo, J. Liu, J. Tang, J. Li, W. Luo, and H. Lu. Aligning linguistic words and visual semantic units for image captioning. 2019. 1

[10] S. Gupta and J. Malik. Visual semantic role labeling. *arXiv preprint arXiv:1505.04474*, 2015. 2, 6

[11] T. Gupta, A. Schwing, and D. Hoiem. No-frills human-object interaction detection: Factorization, layout encodings, and training techniques, 2018. 1

[12] B. Kim, T. Choi, J. Kang, and H. J. Kim. Uniondet: Union-level detector towards real-time human-object interaction detection. In *European Conference on Computer Vision*, 2020. 1, 2, 7, 8, 10, 12

[13] D. Kingma and J. Ba. Adam: A method for stochastic optimization. *International Conference on Learning Representations*, 12 2014. 6

[14] L. Li, Z. Gan, Y. Cheng, and J. Liu. Relation-aware graph attention network for visual question answering. 2019. 1

[15] Y.-L. Li, L. Xu, X. Liu, X. Huang, Y. Xu, M. Chen, Z. Ma, S. Wang, H.-S. Fang, and C. Lu. Hake: Human activity knowledge engine. *arXiv:1904.06539*, 2019. 8

[16] Y.-L. Li, L. Xu, X. Liu, X. Huang, Y. Xu, S. Wang, H.-S. Fang, Z. Ma, M. Chen, and C. Lu. Pastanet: Toward human activity knowledge engine. In *CVPR*, 2020. 2

[17] Y.-L. Li, S. Zhou, X. Huang, L. Xu, Z. Ma, H.-S. Fang, Y. Wang, and C. Lu. Transferable interactiveness knowledge for human-object interaction detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3585–3594, 2019. 1, 2, 7, 10

[18] Y. Liao, S. Liu, F. Wang, Y. Chen, C. Qian, and J. Feng. Ppdm: Parallel point detection and matching for real-time human-object interaction detection. In *CVPR*, 2020. 1, 2, 7, 10

[19] T. Lin, P. Dollár, R. B. Girshick, K. He, B. Hariharan, and S. J. Belongie. Feature pyramid networks for object detection. *CoRR*, abs/1612.03144, 2016. 2

[20] T. Lin, P. Goyal, R. B. Girshick, K. He, and P. Dollár. Focal loss for dense object detection. *CoRR*, abs/1708.02002, 2017. 2, 4, 6, 8

[21] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 6

[22] D. Liu, Z. Zha, H. Zhang, Y. Zhang, and F. Wu. Context-aware visual policy network for sequence-level image captioning. *CoRR*, abs/1808.05864, 2018. 1

[23] Y. Liu, J. Yuan, and C. W. Chen. ConsNet: Learning Consistency Graph for Zero-Shot Human-Object Interaction Detection. *arXiv e-prints*, page arXiv:2008.06254, Aug. 2020. 7

[24] C. Lu, R. Krishna, M. Bernstein, and L. Fei-Fei. Visual relationship detection with language priors. In *European conference on computer vision*, pages 852–869. Springer, 2016. 2

[25] W. Norcliffe-Brown, E. Vafeias, and S. Parisot. Learning conditioned graph structures for interpretable visual question answering. *CoRR*, abs/1806.07243, 2018. 1

[26] S. Qi, W. Wang, B. Jia, J. Shen, and S.-C. Zhu. Learning human-object interactions by graph parsing neural networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 401–417, 2018. 2

[27] S. Ren, K. He, R. B. Girshick, and J. Sun. Faster R-CNN: towards real-time object detection with region proposal networks. *CoRR*, abs/1506.01497, 2015. 2

[28] D. Shao, Y. Zhao, B. Dai, and D. Lin. Finegym: A hierarchical video dataset for fine-grained action understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 1

[29] X. Sun, X. Hu, T. Ren, and G. Wu. Human object interaction detection via multi-level conditioned network. In *Proceedings of the 2020 International Conference on Multimedia Retrieval*, ICMR '20, page 26–34. Association for Computing Machinery, 2020. 7

[30] M. Tan, R. Pang, and Q. V. Le. Efficientdet: Scalable and efficient object detection. 2019. 6, 8

[31] M. D. e. Tiancai Wang, Tong Yang. Learning human-object interaction detection using interaction points. 2020. 1, 2, 7, 10

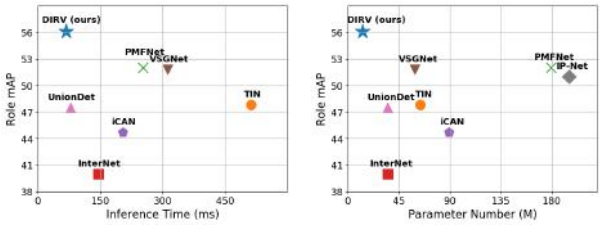
[32] O. Ulutan, A. S. M. Iftekhar, and B. S. Manjunath. Vsgnet: Spatial attention network for detecting human object interactions using graph convolutions. *arXiv*, 2020. 7, 10

[33] B. Wan, D. Zhou, Y. Liu, R. Li, and X. He. Pose-aware multi-level feature network for human object interaction de-

tection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 9469–9478, 2019. 2, 7, 10

[34] J. Wu, L. Wang, L. Wang, J. Guo, and G. Wu. Learning actor relation graphs for group activity recognition. In *CVPR*, 2019. 1

[35] T. Zhou, W. Wang, S. Qi, H. Ling, and J. Shen. Cascaded human-object interaction recognition. pages 4263–4272, 2020. 2, 7



(a) **accuracy** *vs* **speed** (b) **accuracy** *vs* **size**

Figure 5. **mAP versus Inference Time/Parameter Number on V-COCO dataset:** Our proposed DIRV reaches a new state-of-the-art 56.1 mAP_{role} with fastest inference time (68 ms) and fewest parameters (13M) compared with previous methods.

In this supplement, we provide more analysis and experiments not included in the main paper due to space limitation. They are listed as follows:

- Analysis of performance and efficiency is given in Sec. A. We compare our method with other existing ones.
- We show some qualitative results of our proposed *interaction regions* in Sec. B
- More ablation studies are conducted to examine some components of our DIRV in Sec. C.
- We visualize some examples of HOI detection in various cases to analyze the generality of our DIRV in Sec. D.

A. Performance and Efficiency

As mentioned in the main paper, our DIRV surpasses other state-of-the-art approaches in accuracy with both fewer parameters and faster inference speed.

For parameter counting, we follow the estimation strategy in [1] to calculate the parameter number of iCAN [7] and TIN [17]. Similar estimation is also applied to UnionDet [12], InteractNet [8] and IP-Net [31] since the authors did not provide open-source codes. For VSGNet [32] and PMFNet [33], parameters are counted based on the open-source codes.

For time estimation, we consider the sum of the object detection time and HOI detection time for those two-stage approaches, including iCAN [7], TIN [17], InteractNet [8] and VSGNet [32]. We run different models on a NVIDIA RTX2080Ti GPU and some results are referred from other published work [12, 18].

In Fig. 5, we illustrate the performance of different models versus inference time and parameter number separately on V-COCO dataset. It is apparent that our DIRV outperforms others remarkably with a significant superiority in both time and space efficiency.

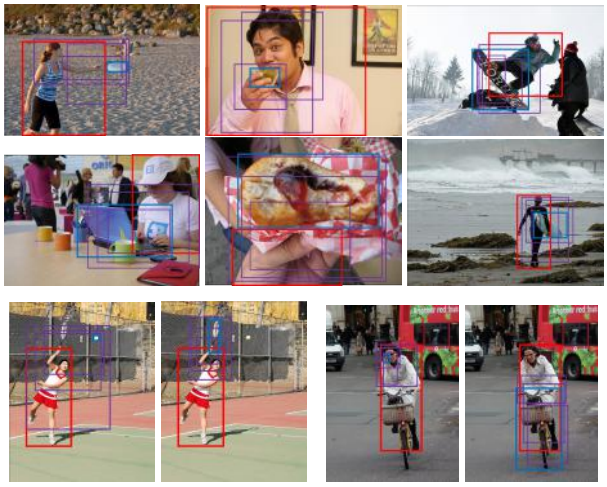


Figure 6. **Examples of Interaction Regions:** *Red* and *blue* rectangles respectively denote the interacting humans and objects. Interaction regions are drawn in *purple*. During training, interaction regions are actually much denser than these illustrations.

B. Analysis of Interaction Regions

We provide more examples of interaction regions with different scales in Fig. 6, where each interaction region is associated with a specific human-object interaction. These interaction regions are composed of parts of the human, object and context, containing visual features essential for HOI detection.

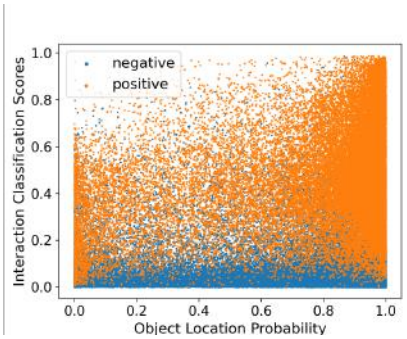


Figure 7. **Scores of Interaction Regions:** It shows *object location probability* $p_{a_j}(x_o^{a_j}, y_o^{a_j})$ and *interaction classification scores* $s_{a_j}^{inter}$ of interaction regions for different human-object pairs. Interaction regions of positive and negative pairs are marked in *orange and blue* separately.

Fig. 7 visualizes the *interaction classification scores* $s_{a_j}^{inter}$ versus *object location probability* $p_{a_j}(x_o^{a_j}, y_o^{a_j})$ of different interaction regions for some human-object pairs. We mark positive (with interactions) and negative (w/o interactions) pairs with different colors. There are two hints in this image. Firstly, positive pairs are predicted with notably higher interaction classification scores in most interaction regions since our interaction regions capture the most crucial visual features for interactions. Secondly, object location probability for positive regions is not certainly very high. The relative spatial relationship is reflected from some very subtle visual features, which are hard to be completely discovered from a single interaction region. This corroborates the necessity of our voting strategy.

As a supplement, we also illustrate *interaction detector* results of some single interaction regions in Fig. 8. Most single regions can fetch satisfying classification results but there exist clear errors in object localization. However, since the errors are distributed in all directions uniformly, they are counterbalanced through voting.

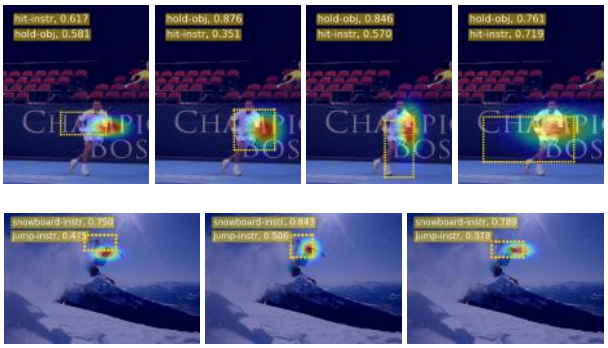


Figure 8. **Detection Result of Single Interaction Region:** *Yellow dotted* rectangles denote the interaction regions. We visualize the object location distribution and list the classification results for each interaction region. Note that here *hit-instr* means *hit with instrument* in the upper line, whose target object is the racket.

C. More Ablation Studies

We add three extra ablation studies on V-COCO dataset here. Firstly, we verify the significance of *interaction detector*, which serves as the key of our DIRV. Then, we examine the effect of backbone networks. Eventually, we consider different values for the standard deviation σ of the 2-d Gaussian distribution in Eq. 8 of the main paper.

C.1. Interaction Detector

Since the results can be derived from the *instance detector* alone, we try to eliminate the whole *interaction detector*. In this case, we have

$$S_{h,o} = s_h \cdot s_o \cdot (s_h^{act} + s_o^{act}) \tag{16}$$

in place of Eq. 14 in the main paper. Results in Tab. 6 witness a dramatic drop of 15.5 mAP, which verifies the indispensability of our novel *interaction detector*.

C.2. Backbone

In our main paper, ablation study in Sec. 4.4 has examined the significance of our dense interaction regions,

Table 6. **Results with Instance Detector Alone:** The lack of interaction detector brings significant performance drop. And EfficientDet backbone only leads to limited improvement compared to ResNet-50.

Method	**Backbone**	**mAP**$_{role}$
Instance Detector	ResNet-50-FPN	38.4
Instance Detector	EfficientDet-d3	40.6
DIRV	EfficientDet-d3	**56.1**

voting strategy and ignorance loss separately in three subsections. Further, we want to ensure that our improvement doesn't come from the backbone solely. In Tab. 6, we compare the performance of two baselines with only the instance detector. They are equipped with two different backbones: ResNet-50-FPN and EfficientDet-d3. The former result comes from this paper [12]. It is noticeable that the EfficientDet backbone only brings a modest elevation compared to the common used ResNet-50-FPN.

C.3. Standard Deviation for Location Distribution

In this part, we analyze the hyper-parameter σ for the Gaussian distribution of the relative object location (Eq. 8 in the main paper). We find that the model performance is not very sensitive to this standard deviation, as is shown in Fig. 9. It shows the reliability and robustness of our interaction region prediction and voting strategy.

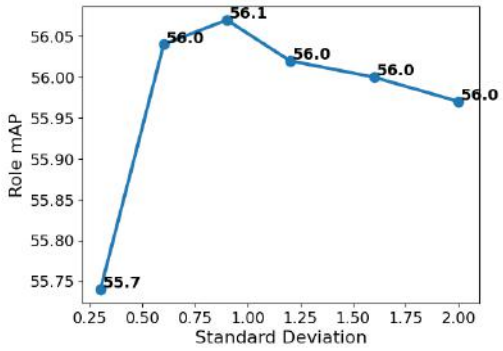


Figure 9. **Ablation Study for Standard Deviation σ of Object Location Distribution**

D. Detection Visualization

We present some visualization of DIRV results in Fig. 10. For each human, the corresponding interaction labels and target objects are displayed in the same color. We mainly pay attention to examples with different characteristics. In these examples, our proposed DIRV deals with various situations very well, despite their special difficulties as follows.

For humans and objects vary in different sizes, our dense interaction regions can easily capture visual features of different scales, resulting in high confidence and accuracy.

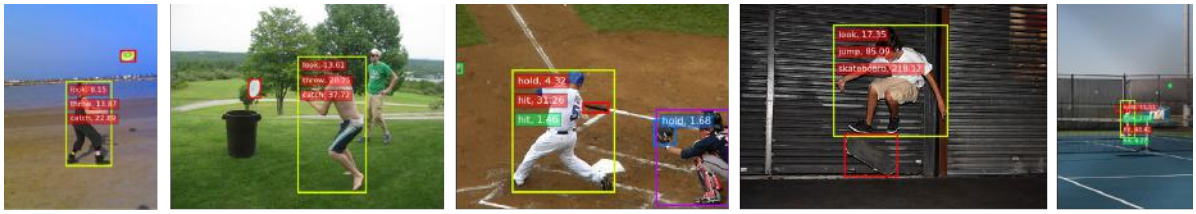
For objects remote from humans, there exist less interactive clues in interaction regions, which makes the prediction harder than close human-object pairs. Despite the overall great performance, several ambiguous interactions (*e.g.* catch and throw) share some common features, bringing possible detection flaws. Multiple interactions of same or different humans may share overlapping interaction regions, which generates potential confusion during training. Yet, our proposed DIRV solved these problems well, obtaining satisfying performance in these cases.

Since our interaction regions focus on parts of humans or objects most essential for interaction, incomplete human or object instances can hardly have any negative influence on the detection.

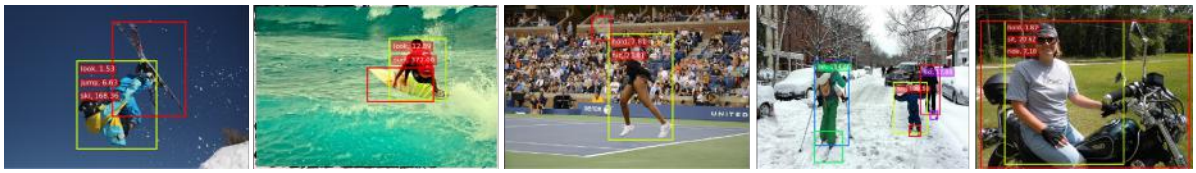
All the examples above verify the strong generality of our proposed approach. We are looking forward to its wide application in different practical applications.

(a) large humans or objects

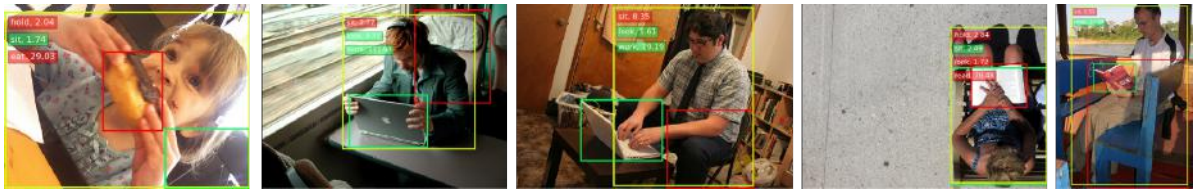
(b) small humans or objects



(c) humans remote from target objects



(d) humans close to target objects



(e) humans interacting with multiple objects

(f) different HOI pairs overlapping with each other

(g) incomplete humans or objects

Figure 10. **Visualization of Detection Results**