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 is publicly available at www.github.com/MVIG-SJTU/DIRV.

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 (Guo et al. 2019; Liu et al. 2018), visual question answering (Li et al. 2019b; Norcliffe-Brown, Vafeias, and Parisot 2018) and action understanding (Pang et al. 2020; Shao et al. 2020).

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 (Gupta, Schwing, and Hoiem 2019; Gao, Zou, and Huang 2018; Li et al. 2019d), while recent one-stage methods aim to regress this union region with keypoints (Liao et al. 2020; Wang et al. 2020) or anchor boxes (Kim et al. 2020) and use it to associate the target human and object.

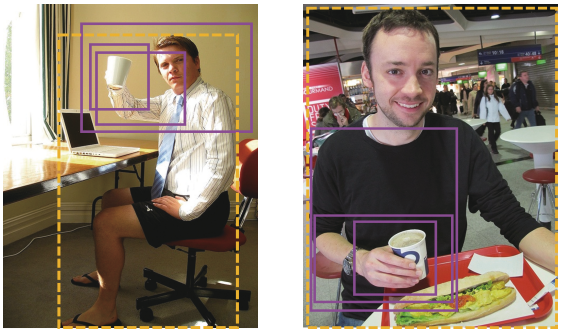


Figure 1: Union Regions vs Interaction Regions: Conventional approaches usually pays attention to the union region (dashed yellow), which contains too much redundant information. Instead, we propose a method focusing on interaction regions (solid violet) with different scales. In above two figures, despite 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 regions 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 includes the

[*] Equal contribution. Names in alphabetical order.
[†] Cewu Lu is corresponding author, member of Qing Yuan Research Institute and MoE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University, China and Shanghai Qi Zhi institute

minimal area of human and object crucial for recognizing the interaction. An example is given in Fig. 1. In this case, an interaction 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 hypothesize 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 (Li et al. 2020), 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 have 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 (Lin et al. 2017) 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*.

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 (Lin et al. 2016; Ren et al. 2015) 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 (Gupta and Malik 2015) 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 (Chao et al. 2018; Gao, Zou, and Huang 2018; Qi et al. 2018), human pose (Fang et al. 2018b; Li et al. 2019d,a), word embedding (Bansal et al. 2019; Lu et al. 2016), segmentation (Wang et al. 2019; Fang et al. 2018a) and human part label (Li et al. 2020). 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.* (Liao et al. 2020) and Wang *et al.* (Wang et al. 2020) 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.* (Kim et al. 2020) 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, achieving a promising accuracy without the help of any other proposals or annotations. The one-stage strategy and concise network architecture also bring greatly improvement in running time and space efficiency.

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. . Then, we present the network architecture of our detector in Sec. . Afterwards, the inference protocol based on *voting strategy* is shown in Sec. . Finally, we demonstrate how to train our deep neural network model in Sec. .

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 (Fang et al. 2017), conventional two-stage

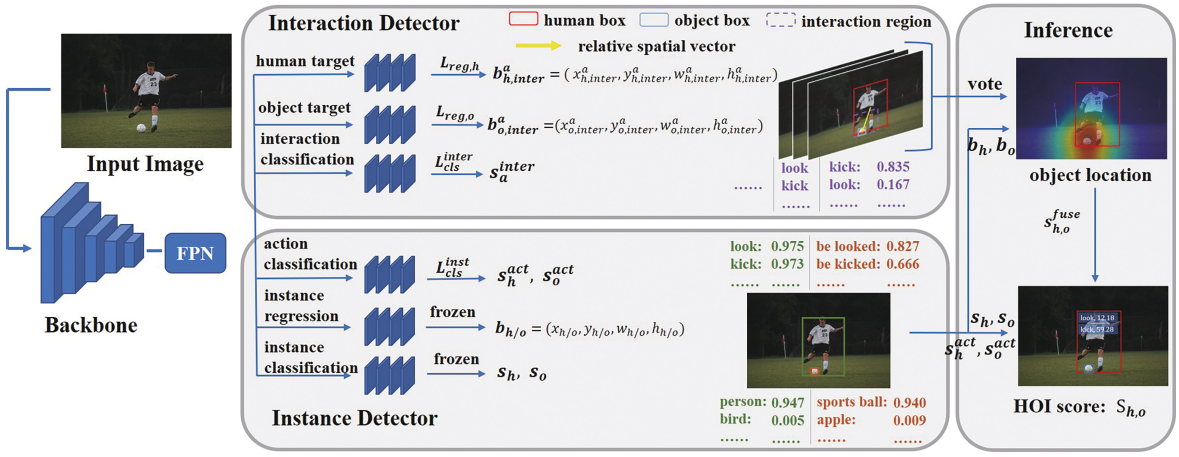


Figure 2: Overview of our DIRV Framework: It is composed of two components: Interaction Detector and Instance Detector. For each interaction region, a relative spatial vector is obtained by regressing the human and object bounding boxes. During inference, results of interaction regions vote for an object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

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. . $vote(\cdot)$ is the voting strategy, which is elaborated in Sec. .

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 (Lin et al. 2016). We first explain the instance detector for $\mathcal{H}, \mathcal{O}$ and then the interaction detector for $\langle T(b_h^i), v^i, T(b_o^i) \rangle$.

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 classification. 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.

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, 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..

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:

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

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 (Lin et al. 2017) 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.

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.

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*). 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.

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 (Gkioxari et al. 2018), we model it with a 2-d Gaussian distribution:

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

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. As analyzed in the supplementary material, our method is insensitive to σ.

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.

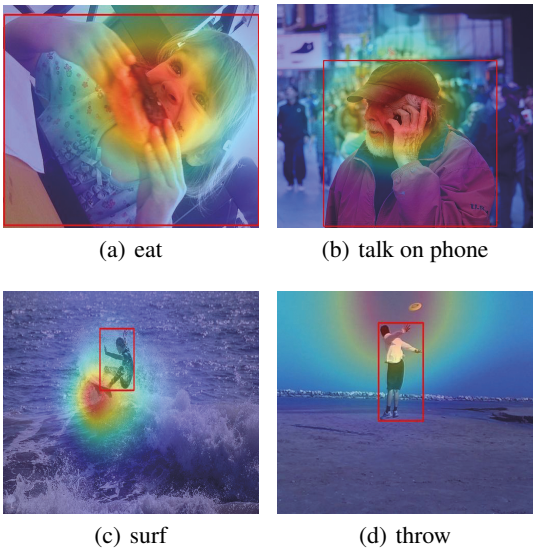
Voting Based Region Fusion By fusing *weighted localization scores* of interaction regions associated with same human instance b_h, a *human-centric object location distribution* s_h^{fuse} is computed with our voting strategy:

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

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.

Finally, 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.

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



(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.

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.

Model Training

During training, the backbone, feature pyramid network and instance classification/regression branches are frozen with COCO pre-trained weight (Tan, Pang, and Le 2020). 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 (Lin et al. 2017) 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*.

Experiments

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

Dataset and Metric

Dataset We evaluate our method on two popular datasets: **V-COCO** (Gupta and Malik 2015) and **HICO-DET** (Chao et al. 2015). V-COCO dataset is a subset of COCO (Lin et al. 2014) 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 (Chao et al. 2015) 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.

Implementation Details

For HOI detection, we use EfficientDet-d3 (Tan, Pang, and Le 2020) 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 (Kingma and Ba 2014) 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.

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 (top columns in both Tab. 1,2) into account in our comparison. What's more, unlike many existing two-stage approaches, our method does not rely on object proposals, which significantly elevates its compatibility.

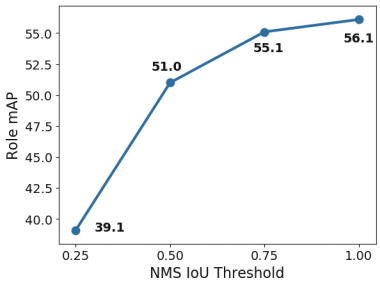


Figure 4: Ablation Study for Voting Strategy: The mAP_{role} increases as the IoU threshold for NMS grows. There is actually no NMS when IoU threshold is 1.

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 (Tab. 2), we report the results on two different settings: Default and Known Objects. The interaction classification branch only classifies verb categories *e.g. eating*, which are associated with object categories *e.g. apple* based on the results of instance classification branch, as in (Kim et al. 2020). This classification strategy brings a more promising performance than directly recognizing the verb-object pair. The reason may be that it reduces the number of categories in interaction classification branch, which elevates the accuracy. What's more, it also saves the space overhead, allowing a larger batch size during training and improving the training stability. The results also demonstrated that our approach has a superiority in time and space complexity.

Two prior arts share some common insights with us. InteractNet (Gkioxari et al. 2018) localizes objects based on single human appearance. UnionDet (Kim et al. 2020) is another anchor-based one-stage HOI detection approach, focusing on *union regions*. However, 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.

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.

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.

Method	Prop.	Ext.	mAP$_{role}$	Time	Params
RP_DC_D (Li et al. 2019d)	✓	P	47.8	513 ms	64 M
PMFNet (Wan et al. 2019)	✓	P	52.0	253 ms	179 M
ConsNet (Liu, Yuan, and Chen 2020)	✓	P+L	53.2	-	-
MLCNet (Sun et al. 2020)	✓	P+B+L	55.2	-	-
InteractNet (Gkioxari et al. 2018)	✓	×	40.0	145 ms	71 M
Zhou *et al.* (Zhou et al. 2020)	✓	×	48.9	-	620 M
VSGNet (Ulutan, Iftekhar, and Manjunath 2020)	✓	×	51.8	312 ms	59 M
UnionDet (Kim et al. 2020)	×	×	47.5	78 ms	44 M
IP-Net (Wang et al. 2020)	×	×	51.0	-	-
DIRV (ours)	×	×	**56.1**	**68 ms**	**12 M**

Table 1: Results on V-COCO: Prop. shows whether it needs object detection beforehand. Ext. means extra features, where P,B,L denotes human pose, human body part states and language priors respectively, which are utilized in prior methods.

Method	Prop.	Ext.	Default			Known Object			Time	Params
			F	R	NR	F	R	NR		
RP_DC_D (Li et al. 2019d)	✓	P	17.03	13.42	18.11	19.17	15.51	20.26	513 ms	64 M
PMFNet (Wan et al. 2019)	✓	P	17.46	15.65	18.00	20.34	17.47	21.20	253 ms	179 M
MLCNet (Sun et al. 2020)	✓	P+B+L	17.95	16.62	18.35	22.28	20.73	22.74	-	-
ConsNet (Liu, Yuan, and Chen 2020)	✓	P+L	22.15	17.12	23.65	-	-	-	-	-
InteractNet (Gkioxari et al. 2018)	✓	×	9.94	7.16	10.77	-	-	-	145 ms	72 M
UnionDet (Kim et al. 2020)	×	×	17.58	11.72	19.33	19.76	14.68	21.27	78 ms	50 M
IP-Net (Wang et al. 2020)	×	×	19.56	12.79	21.58	22.05	15.77	23.92	-	-
PPDM-Hourglass (Liao et al. 2020)	×	×	21.73	13.78	**24.10**	24.58	16.65	26.84	71 ms	195 M
DIRV (ours)	×	×	**21.81**	**16.35**	23.44	**25.84**	**21.02**	**27.28**	**68 ms**	**13 M**

Table 2: Results on HICO-DET: Prop. shows whether it needs object detection beforehand. Ext. means extra features, where P,B,L denotes human pose, human body part states and language priors respectively, which are utilized in prior methods. The columns F, R, NR denote Full, Rare, Non-Rare separately.

$\mathbf{t_h}$	$\mathbf{t_o}$	$\mathbf{t_u}$	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**

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.

Loss Function	mAP$_{role}$
Focal Loss (Lin et al. 2017)	54.8
Foreground Loss (Kim et al. 2020)	54.0
Ignore Loss (ours)	**56.1**

Table 4: Loss Function for Interaction Classification

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 confidence may not be very high.

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 (Kim et al. 2020) and our proposed *ignorance loss*. Results in Tab. 4 verify our superiority since it can help to deal with region overlapping and missed positive labels.

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 incorporate the part-level knowledge (Li et al. 2019c) into our framework.

Acknowledgements

This work is supported in part by the National Key R&D Program of China, No. 2017YFA0700800, National Natural Science Foundation of China under Grants 61772332, Shanghai Qi Zhi Institute, SHEITC(2018-RGZN-02046) and Baidu Fellowship.

References

Bansal, A.; Rambhatla, S. S.; Shrivastava, A.; and Chellappa, R. 2019. Detecting Human-Object Interactions via Functional Generalization. *CoRR* abs/1904.03181.

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

Chao, Y.-W.; Wang, Z.; He, Y.; Wang, J.; and Deng, J. 2015. HICO: A Benchmark for Recognizing Human-Object Interactions in Images. In *Proceedings of the IEEE International Conference on Computer Vision*.

Fang, H.; Lu, G.; Fang, X.; Xie, J.; Tai, Y.; and Lu, C. 2018a. 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)*, 70–78.

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

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

Gao, C.; Zou, Y.; and Huang, J.-B. 2018. ican: Instance-centric attention network for human-object interaction detection. *arXiv preprint arXiv:1808.10437* .

Gkioxari, G.; Girshick, R.; Dollár, P.; and He, K. 2018. Detecting and Recognizing Human-Object Interactions. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Guo, L.; Liu, J.; Tang, J.; Li, J.; Luo, W.; and Lu, H. 2019. Aligning linguistic words and visual semantic units for image captioning. In *Proceedings of the 27th ACM International Conference on Multimedia*, 765–773.

Gupta, S.; and Malik, J. 2015. Visual semantic role labeling. *arXiv preprint arXiv:1505.04474* .

Gupta, T.; Schwing, A.; and Hoiem, D. 2019. No-frills human-object interaction detection: Factorization, layout encodings, and training techniques. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9677–9685.

Kim, B.; Choi, T.; Kang, J.; and Kim, H. J. 2020. UnionDet: Union-level Detector Towards Real-Time Human-Object Interaction Detection. In *European Conference on Computer Vision*.

Kingma, D.; and Ba, J. 2014. Adam: A Method for Stochastic Optimization. *International Conference on Learning Representations* .

Li, J.; Wang, C.; Zhu, H.; Mao, Y.; Fang, H.-S.; and Lu, C. 2019a. Crowdpose: Efficient crowded scenes pose estimation and a new benchmark. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 10863–10872.

Li, L.; Gan, Z.; Cheng, Y.; and Liu, J. 2019b. Relation-aware graph attention network for visual question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10313–10322.

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

Li, Y.-L.; Xu, L.; Liu, X.; Huang, X.; Xu, Y.; Wang, S.; Fang, H.-S.; Ma, Z.; Chen, M.; and Lu, C. 2020. PaStaNet: Toward Human Activity Knowledge Engine. In *CVPR*.

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

Liao, Y.; Liu, S.; Wang, F.; Chen, Y.; Qian, C.; and Feng, J. 2020. PPDM: Parallel Point Detection and Matching for Real-time Human-Object Interaction Detection. In *CVPR*.

Lin, T.; Dollár, P.; Girshick, R. B.; He, K.; Hariharan, B.; and Belongie, S. J. 2016. Feature Pyramid Networks for Object Detection. *CoRR* abs/1612.03144.

Lin, T.; Goyal, P.; Girshick, R. B.; He, K.; and Dollár, P. 2017. Focal Loss for Dense Object Detection. *CoRR* abs/1708.02002.

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

Liu, D.; Zha, Z.; Zhang, H.; Zhang, Y.; and Wu, F. 2018. Context-Aware Visual Policy Network for Sequence-Level Image Captioning. *CoRR* abs/1808.05864.

Liu, Y.; Yuan, J.; and Chen, C. W. 2020. ConsNet: Learning Consistency Graph for Zero-Shot Human-Object Interaction Detection. In *Proceedings of the 28th ACM International Conference on Multimedia*, 4235–4243.

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

Norcliffe-Brown, W.; Vafeias, E.; and Parisot, S. 2018. Learning conditioned graph structures for interpretable visual question answering. *arXiv preprint arXiv:1806.07243* .

Pang, B.; Zha, K.; Zhang, Y.; and Lu, C. 2020. Further Understanding Videos through Adverbs: A New Video Task. In *AAAI*, 11823–11830.

Qi, S.; Wang, W.; Jia, B.; Shen, J.; and Zhu, S.-C. 2018. Learning human-object interactions by graph parsing neural

networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 401–417.

Ren, S.; He, K.; Girshick, R. B.; and Sun, J. 2015. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *CoRR* abs/1506.01497.

Shao, D.; Zhao, Y.; Dai, B.; and Lin, D. 2020. FineGym: A Hierarchical Video Dataset for Fine-Grained Action Understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Sun, X.; Hu, X.; Ren, T.; and Wu, G. 2020. Human Object Interaction Detection via Multi-Level Conditioned Network. In *Proceedings of the 2020 International Conference on Multimedia Retrieval*, ICMR '20, 26–34. Association for Computing Machinery. ISBN 9781450370875. URL https://doi.org/10.1145/3372278.3390671.

Tan, M.; Pang, R.; and Le, Q. V. 2020. Efficientdet: Scalable and efficient object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10781–10790.

Ulutan, O.; Iftekhar, A.; and Manjunath, B. S. 2020. Vsgnet: Spatial attention network for detecting human object interactions using graph convolutions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13617–13626.

Wan, B.; Zhou, D.; Liu, Y.; Li, R.; and He, X. 2019. Pose-aware Multi-level Feature Network for Human Object Interaction Detection. In *Proceedings of the IEEE International Conference on Computer Vision*, 9469–9478.

Wang, T.; Yang, T.; Danelljan, M.; Khan, F. S.; Zhang, X.; and Sun, J. 2020. Learning human-object interaction detection using interaction points. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4116–4125.

Wang, W.; Zhang, Z.; Qi, S.; Shen, J.; Pang, Y.; and Shao, L. 2019. Learning compositional neural information fusion for human parsing. In *Proceedings of the IEEE International Conference on Computer Vision*, 5703–5713.

Zhou, T.; Wang, W.; Qi, S.; Ling, H.; and Shen, J. 2020. Cascaded human-object interaction recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4263–4272.