Image and Vision Computing 121 (2022) 104422

Contents lists available at ScienceDirect

Image and Vision Computing

journal homepage: www.elsevier.com/locate/imavis

Effective actor-centric human-object interaction detection

Kunlun Xu, Zhimin Li, Zhijun Zhang, Leizhen Dong, Wenhui Xu, Luxin Yan, Sheng Zhong, Xu Zou *

School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, Wuhan 430074, China

ARTICLE INFO

Article history:
Received 25 December 2021
Accepted 19 February 2022
Available online 25 February 2022

Keywords:
Human-object interaction detection
Global context utilizing
Pixel-wise prediction
Deep learning

ABSTRACT

While Human-Object Interaction (HOI) Detection has achieved tremendous advances in recent, it still remains challenging due to complex interactions with multiple humans and objects occurring in images, which would inevitably lead to ambiguities. Most existing methods either generate all human-object pair candidates and infer their relationships by cropped local features successively in a two-stage manner, or directly predict interaction points in a one-stage procedure. However, the lack of spatial configurations or reasoning steps of two- or one-stage methods respectively limits their performance in such complex scenes. To avoid this ambiguity, we propose a novel actor-centric framework. The main ideas are that when inferring interactions: 1) the non-local features of the entire image guided by actor position are obtained to model the relationship between the actor and context, and then 2) we use an object branch to generate pixel-wise interaction area prediction, where the interaction area denotes the object central area. Moreover, we also use an actor branch to get interaction prediction of the actor and propose a novel composition strategy based on center-point indexing to generate the final HOI prediction. Thanks to the usage of the non-local features and the partly-coupled property of the human-objects composition strategy, our proposed framework can detect HOI more accurately especially for complex images. Extensive experimental results show that our method achieves the state-of-the-art on the challenging V-COCO and HICO-DET benchmarks and is more robust especially in multiple persons and/or objects scenes.

1. Introduction

Human-Object Interaction (HOI) detection [1–3], aiming to detect human, object and corresponding inter-action between them from a given image, is a meaningful task serving as the fundamental step for many computer vision applications such as robotics [4–6] and activity analysis [7]. It is still a challenging task due to the inevitable ambiguities that come from the various visual relationships between multiple persons and objects occur-ring in the scene.

Given an image, most existing methods [8–12] detect humans and objects first and, for each human-object pair, recognize the probable interactions with the local features cropped according to their bounding boxes. However, though such a pipeline works well for simple scenes when only appears few people, it may fail on com-plex scenes with ambiguities caused by multiple persons and objects, as shown in Fig. 1. To tackle this issue, a few approaches [13–15] try to utilize context information to complement local appearance to obtain a better detection result. Some of them [13,14,16] seek to ex-plore contextual representation as supportive information for the cropped local features, and improve detection re-sults stably. However, they usually design an additional branch to extract contextual information by Global Av-erage Pooling (GAP), which would cause an information loss, especially spatial configurations due to the pooling operation of all pixels. Thus such utilization of contextual information is limited. Other crop-free approaches [15,17] formulate HOI detection as an interaction point detection problem and use one-stage strategy to exploit contextual information directly. Although these methods are quite effective in many cases, their highly-coupled property and the lack of reasoning steps [18] make the model hard to work well on complex images.

To overcome these issues, we present a novel partly-coupled actor-centric HOI detection framework. On the one hand, we formulate HOI detection task as a pixel-wise prediction problem in a one-stage like procedure, which avoids the information loss by GAP operations. On the other hand, non-local features guided by binary masks generated from each detected human without the crop-ping procedure would be obtained. On the basis of these non-local features, every single human is able to evaluate the existence of interaction for all objects in the image. More specifically, we present an actor-centric model. In this model, we propose RGBM Generator module, which generates binary mask (M) according to human bounding box and concatenate it with the original image (RGB). The generated RGBM data is used as input of feature extraction backbone. Then, we regress interacting confidence map explicitly with a square mask supervision which could be inferred by original annotations directly. Finally, we propose an efficient post-

* Corresponding author.
E-mail address: zoux@hust.edu.cn (X. Zou).

https://doi.org/10.1016/j.imavis.2022.104422
0262-8856/

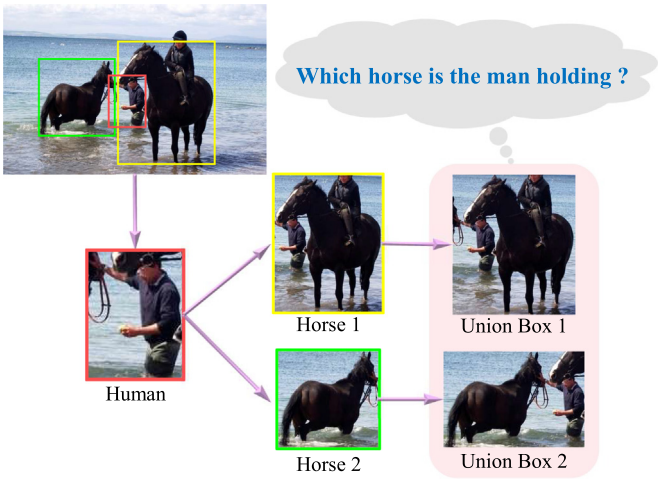


Fig. 1. How to differentiate which horse is interacting with the person in the red box? The horse 1 (in the yellow box) is very close to the human, resulting in ambiguities between two interaction pairs with only cropped local features [8,9]. This motivates us to consider multiple objects in the image simultaneously while predicting one person's interactions to avoid this ambiguity.

processing procedure to generate HOI predictions for this brand-new framework, in which a composition strategy based on center-point indexing is used. Thanks to the usage of the non-local features and the partly-coupled property of the human-objects composition strategy, our proposed framework can detect HOI more accurately especially for complex images. To summarize, our contributions are four-folded:

- We propose a novel actor-centric HOI detection framework to explore the relationships between one human and multiple objects, addressing the ambig-uity issue across multiple interactions in the image.
- We formulate the HOI detection task as a pixel-wise classification problem in a one-stage-like procedure with a proposed Weighted Cross Entropy Loss (WCEL), which could reduce the overfit of the boundary hop.
- We present an RGBM Generator module to provide an actor-centric guidance of the network, and design an efficient composition strategy to obtain the final scores by a scoremap indexing post-process.
- Extensive experimental results and ablation studies demonstrate the superiority of our approach in both quantitative and qualitative results in challenging HICO-DET and V-COCO benchmarks, especially for complex images.

2. Related work

Existing HOI detection methods can be roughly categorized into crop-based and crop-free fashions [18–21].

Crop-based HOI detection methods [8–13] utilize Faster R-CNN or Mask R-CNN to generate human /object bounding box with the corresponding confidence in the first stage. And in the second stage, the interaction between human and object is detected by parsing the cropped features in the union box. Chao et al. [10] pro-pose a multi-stream framework including human-stream, object-stream and pairwise-stream to obtain verb score, and then the verb confidence is obtained by adding three scores from all streams. Gao et al. [13,16] enhance the feature in human/object stream through attention mechanism and obtain a significant improvement thanks to aggregating the global information.

However, their performance is limited because the spatial and contextual information are often lost during the cropping procedure, which might cause ambiguous situations, as shown in Fig. 1. In contrast, our method recognizes interaction in a crop-free manner. We leverage a global context extraction network instead of cropped feature fusion network to make sure every pixel is able to obtain global information. With the supervision of our proposed loss, each pixel is guided to output interaction score of the object/human it belongs to.

Recently, crop-free HOI detection methods [15,17] have attracted increasing attention because previous works struggle with the problem of efficiency. Most crop-free methods extract global-aware features for better perform-ance and take HOI as a key-point detection task. IP-Net [17] detects the interactions between human-object pairs through Hourglass-104 [22,23] and then associates human and object proposal generated by FPN [24] to obtain final HOI predictions. PPDM [15] defines the localization for interaction. Human and object points are the center of the detection boxes, and the interaction point is the midpoint of the human and object points. It utilizes a matching strategy in the post-processing, and makes the pipeline in one-stage.

Although the efficiency of one-stage methods is usually relatively satisfactory, the performance suffered since there might exist conflicts of interaction keypoints when multiple persons or objects are shown. To avoid such conflicts, we innovatively constrain the network to predict the interaction at the central area of human/object.

3. Methodology

Given an input image $\in R^{W\times H}$, the goal of HOI detection is detecting the ⟨human, verb, object⟩ triplets, such as ⟨human, hold, cup⟩, ⟨human, ride, horse⟩. In this work, we denote the verb set as $\Psi = \{\psi_i\}_{i=1}^{K}$, where K is the total number of verb categories.

3.1. Overview

Our proposed framework is built on a pre-trained object detector and focuses on inferring the triplet categories for each human in the image. As illustrated in Fig. 2, we firstly use an object detector to locate humans and objects, then traverse all detected humans by an Actor Switch module sequentially. The selected human is defined as actor and utilized to generate an RGBM four-channel data (described in the following chapter). Then, a backbone [25,26] is used to extract features, behind which the framework separates into two branches, namely Actor Branch and Object Branch. Actor Branch aims to predict pixel-wise verb scores of the current actor. When the actor is performing one interaction, the corresponding verb scores in the actor's central area is expected to be high. Similarly, for Object Branch, we also obtain per-pixel object verb predictions. For the expectation of the network to focus on the center location of these agents, we narrow the original actor/object box into a smaller binary square mask, which has been shown by the red area in Fig. 2. These red areas are generated as the targets of our network, thus we can cooperate the current actor verb score and the corresponding multiple objects verb scores from the two branches to predict the interaction triplets.

3.2. RGBM generation and feature extraction

We denote detected human set as $H = \{h_i\}_{i=1}^{M}$, where h_i is a detected human and M is the number of detected humans. Similarly, when N objects is detected, we denote detected object set as $O = \{o_i\}_{i=1}^{N}$, where o_i is a detected object.

There are many ways to indicate human position. We choose to use an additional $W \times H$mask map to avoid damaging the original image structure of the RGB channel. As is shown in Fig. 3, given a human box and an input image I, we generate a human position mask map with the same height and width with I. The area within the human box is set as 1 and other positions are set as 0. Then we concatenate the human position mask and origin image. We denote the generated 4-channel data as RGBM image.

In order to ensure every position on the image could utilize global context and local appearance to generate verb prediction, we adopt

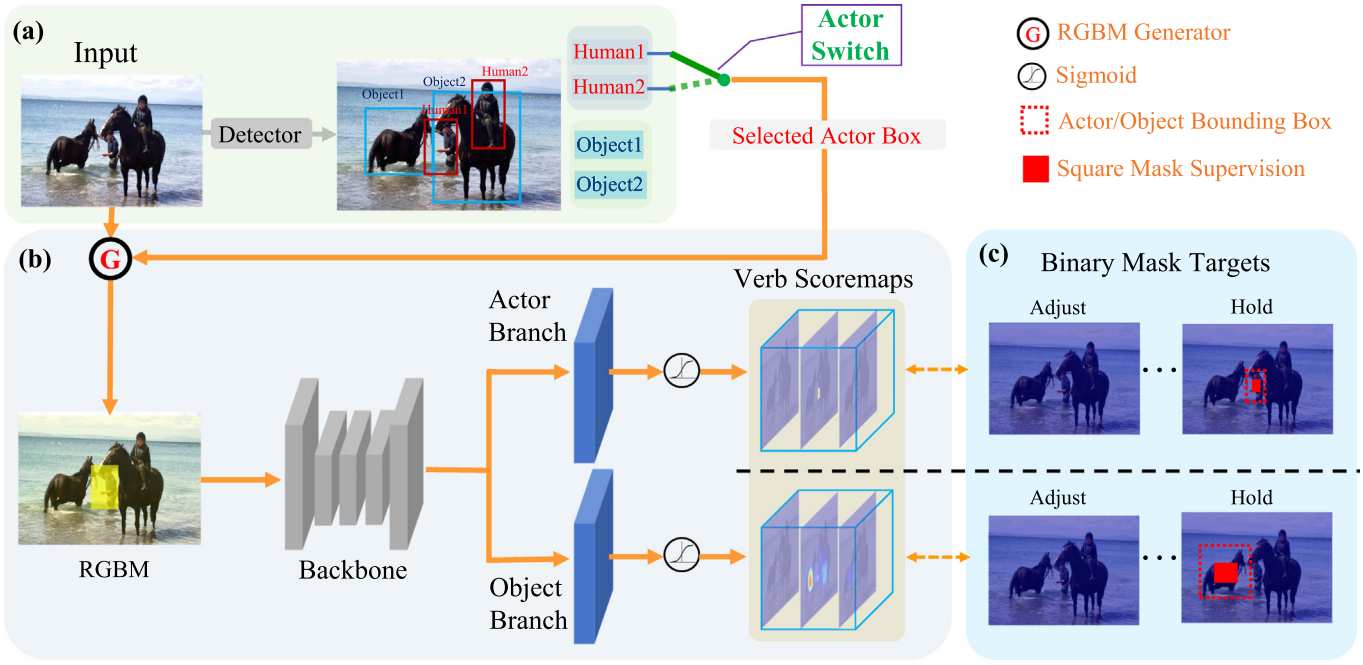


Fig. 2. An overview of our framework for Human-Object Interaction Detection. We formulate the HOI detection task as a pixel-wise verb score prediction problem in two branches, *i.e.*, Actor Branch and Object Branch. The input of the network is generated by an RGBM Generator based on the Actor Switch that selects each human in the image sequentially. In order to force the network to predict interaction location and scores simultaneously, we design a strategy that generates square masks according to original annotations and adopt the masks as supervision. Guided by actor mask in the RGBM data, our framework could exploit the relationship between the actor and multiple objects in the image to generate accurate predictions.

DLA [25] or HRNet [26] which are proposed to effectively aggregate global and local in-formation, as our backbone. After feature extraction, the feature resolution is down from $W \times H$ to $W/d \times H/d$, where d is the down-sample stride.

3.3. Actor branch and object branch

The Actor Branch and Object Branch are two detec-tion heads to generate pixel-wise verb prediction for actor and all objects separately. In Fig. 2, output size of each branch is $W/d \times H/d \times (K+1)$ where K is the number of annotated verb categories. We use the first K channels to generate predictions for all verb categories. The last channel is used to generate prediction for an additional category which is named w/o-interaction.

For Actor Branch, we request the network to gener-ate a high score in the area around actor box center when an interaction happens. In practice, we use a box smaller than actor box as the central area which is denoted as Z. Furthermore, we denote the interaction set the actor par-ticipates in as Ψ_a. Note that if the actor is not interacting with any object, we set $\Psi_a = \{\text{w/o-interaction}\}$. And we generate target illustrated in Fig. 2, by

$$f_a(x,y,c) = \begin{cases} 1 & (x,y) \in Z,\ \Psi^*[c] \in \Psi_a \\ 0 & otherwise \end{cases} \quad (1)$$

where (x,y) denotes a spatial position and c denotes channel. Ψ^* is expanded from origin verb set Ψ by adding the additional category w/o-interaction. $\Psi^*[c] \in \Psi^*$ is the verb category channel c represents.

For Object Branch, given $o_i \in O$, we denote the inter-action set o_i participates in with the actor as Ψ_{o_i} and we set $\Psi_{o_i} = \{\text{w/o-interaction}\}$ if o_i is not interacting with the actor. We also generate target for each object by Equ. 1 and denote the generated target is f_{o_i}. Note that the w/o-interaction category here means oi is not interacting with the actor. The target for Object Branch denoted as f_o is calculated by

$$f_o = f_{o_1} \oplus f_{o_2} \cdots \oplus f_{o_N} \quad (2)$$

where $\oplus$ denotes an addition rule that retains the maxi-mum for elements in the same position.

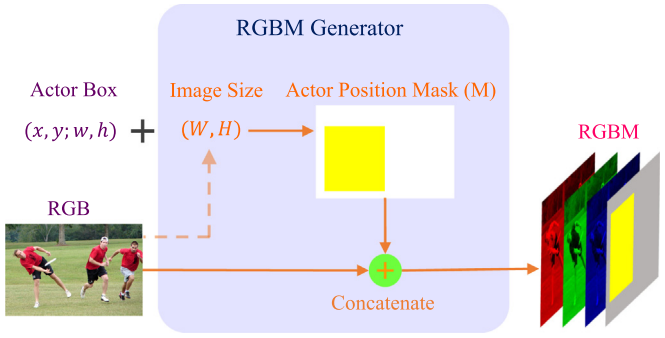


Fig. 3. Illustration of RGBM data generation. We first generate a Actor Position Mask, *i.e.*, $M \in \mathbb{R}^{W \times H \times 1}$, based on the image size of the RGB image and coordinates of the Actor box. Then we concatenate the mask with the original RGB image to obtain four-channel RGBM data as the inputs of our framework.

3.4. Loss

Instead of simply using Cross Entropy Loss, we pro-pose a Weighted Cross Entropy Loss (WCEL) based on following design.

Hanning Weight: Because there is no transition between box central area and other area, the network may pay too much attention on learning a sharp hop at the boundary of box central area and this could cause overfit. To settle this, we want the position near the center point has a higher weight and position near the boundary has a lower weight. In this work, we use Hanning window to generate weight for the positive positions and negative positions individually. The two-dimensional Hanning window is def-ined as:

$$\mathcal{H}(x,y,w,h)=\frac{1}{4}\left[1+\cos\left(2\pi\frac{x}{w-1}\right)\right]\left[1+\cos\left(2\pi\frac{y}{h-1}\right)\right] \tag{3}$$

where w and h denote width and height of Hanning window and (x,y) is a point within Hanning window.

Given a box B whose width, height and center point are w, h and (x_0,y_0) separately, the proposed Hanning Weight w_{han} for positions within B is calculated by

$$w_{han}(x,y,c)=\begin{cases}\mathcal{H}(x-x_0,y-y_0,w,h) & f(x,y,c)=1\\ 1-\mathcal{H}(x-x_0,y-y_0,w,h) & f(x,y,c)=0\end{cases} \tag{4}$$

And we set $w_{han}(x,y,c)=1$ if (x,y) is not within B.

Scale Weight: Because the size of the central area is rel-evant to box size, if we simply use classification loss for every spatial position, bigger box will have bigger area to generate loss and smaller ones has smaller area to generate loss. Such imbalance of supervision area will cause smaller object/actor to become hard to optimize. To solve this problem, we give different area corresponding wights according to box position and size. We define this kind of weight as Scale Weight and express it as w_{scale}. Take the above box B for example, the Scale Weight of position within B is calculated by:

$$w_{scale}(x,y,c)=min(10,\lambda_s max(W,H)/max(w,h)) \tag{5}$$

where λ_s is a super parameter and we set it as 0.5 in our work. Besides, we set the upper limit of w_{scale} as 10 to prevent the weight of small box greater than that of background too much. Similar to Hanning Weight, we set $w_{scale}(x,y,c)=1$ if (x,y) is not within B.

Weighted Cross Entropy Loss: Assume the output of network is $\widehat{f}\in R^{W\times H\times(K+1)}$ and corresponding target, Hanning Weight and Scaled Weight is y, w_{han}, w_{scale} separately. The loss for $\widehat{f}$, denoted as $\mathcal{L}$, is calculated by

$$\begin{aligned}\mathcal{L}=\sum_{x,y,c}w_{han}(x,y,c)w_{scale}(x,y,c)\Big[&f(x,y,c)\,log\widehat{f}(x,y,c)\\&+(1-f(x,y,c))\,log\left(1-\widehat{f}(x,y,c)\right)\Big]\end{aligned} \tag{6}$$

Assume $\widehat{f}_a$ and $\widehat{f}_o$ are outputs of Actor Branch and Object Branch respectively. We denote the loss of Actor Branch and Object Branch as a and o, and they can be calculated by Equ. 6.

Total loss of our framework is denoted as

$$\boldsymbol{L}=\lambda_a\mathcal{L}_a+\lambda_o\mathcal{L}_o \tag{7}$$

where λ_a and λ_o are two super parameters to balance the losses of Object Branch and Human Branch. In our work, we set $\lambda_a=\lambda_o=1$.

3.5. Inference

During inference, we use the pipeline shown in Fig. 2 to generate pixel-wise verb predictions first. Then, we use a simple post-process strategy illustrated in Fig. 4 to generate final HOI scores. Given the selected actor and object set $O=\{o_i\}_{i=1}^N$ provided by detector, we obtain object center point set $C_o=\{c_{oi}\}_{i=1}^N$ and actor center point c_a. Then, we use c_a to index actor verb score S_a^v. Simultaneously, we use c_{oi} to index object verb scores S_{oi}^v for each $o_i\in O$. Next, we calculate verb score $S_{a,\,oi}^v$ for an actor-o_i pair by

$$S_{a,o}^v=S_a^v\cdot S_o^v \tag{8}$$

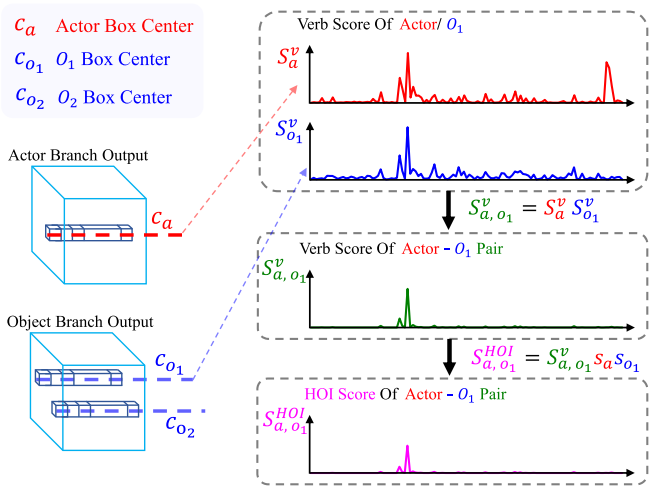


Fig. 4. Illustration of final score predictions. We first calculate verb scores of the actor and objects by indexing on two branches output based on the center point coordinates of their bounding boxes. Then the final HOI prediction can be obtained by multiplying the verb scores and the detection scores.

And the HOI score $S_{a,oi}^{hoi}$ of actor-o_i pair is denoted as

$$S_{a,o_i}^{hoi}=S_{a,o_i}^v\cdot s_a\cdot s_{o_i} \tag{9}$$

where s_{oi} is the object confidence score provided by object detector. s_a is the confidence score of the actor (*i.e.*, selected human) and is also provided by object detector.

4. Experiments

4.1. Datasets and metrics

Datasets. We verify the performance of our method on HICO-DET [10], V-COCO [27] and HOI-A [15]. HICO-DET contains 38,118 images for training and 9658 images for testing. There are 80 object categories as same as MS-COCO [28] and 117 verb categories. The object and verb categories composite 600 HOI categories for HOI detection. V-COCO is a smaller dataset which is derived from MS-COCO dataset. It contains 5400 images in the trainval dataset and 4946 images in the test dataset. Each person has annotations for 29 action categories which contains 25 HOI categories and 4 body motions. HOI-A is a recently proposed dataset containing 11 kinds of objects and 10 action categories in general scenes.

Metrics. Following previous works, we adopt the mean average precision (mAP) to evaluate our method. For HICO-DET, AP is computed on per HOI class and two setting are adopted, *i.e.*, "Known Object" and "Default". In "Known Object" setting, we evaluate only on the images containing the annotated object category for each HOI category, and in "Default" setting, we evaluate on full test images. In each setting, we report the mAP over three types, *i.e.*, Rare, Non-Rare and Full, which are defined according to number of training instances. For V-COCO, three settings are adopted. In AP_{agent}, the true positive only focus on the pair ⟨subject, verb⟩. While AP_{role}^{S1} and AP_{role}^{S2} requires object should be also correctly located. For the cases when there is no object, in AP_{role}^{S1} a prediction is correct if the corresponding bounding box for the object is empty and in AP_{role}^{S2} the bounding box of the object is not considered. For HOI-A, AP is computed on per verb class.

4.2. Implementation details

Object Detector. We use Faster R-CNN [39] as the object detector. In experiments, we adopt the pre-trained model released by MMDetection

[40] to localize persons and objects. The NMS threshold is set as 0.05 and top-100 predictions are used for later processing. We also test our method on HICO-DET by utilizing the detection results released by DRG [35] with the model that finetuned on HICO-DET. When using the detection results of DRG, we set the threshold of both human and object score to 0.2.

Annotation Arrangement. For the convenience of training interaction classification network, we rewrite the training annotation: for each annotated person p, we record box of p, all objects interacting with p and corresponding verbs. For a detected person p_0, if there exists an an-notated person that has IoU over 0.5 with p, we assign object and verb annotation of the person to p_0, otherwise we define p_0 as a negative example. Usually, the negative examples are several times of the positive ones. To tackle the imbalance of examples, we set the ratio between pos-itives and negatives as 1 : 1 by randomly dropping some negative examples.

Feature Extractor. We use HRNet-W32 [16] and DLA-34 [25] as our backbones for interaction classification and evaluating the effectiveness of our proposed idea. HRNet was initially proposed for human pose estimation which requires the network to extract local and global features for effective keypoint detection. The HRNet-W32 used in this work is a relatively lightweight edition compared to HRNet-W48 and HRNet-W64. DLA is a more lightweight network, which is proposed to fuse information across lay-ers of network, and it can also make sure every position in the output layer obtain information across scales. In our experiments, we initialize HRNet-W32 and DLA-34 with weights pretrained on ImageNet [50].

Other Parameters. The bounding box of the supervised area illustrated in Fig. 2 is scaled from human/object box with a ratio of 0.3. During training, we use ADAM [51] as the optimizer and train the model for 12 epochs with learning rate 1.5×10^{-5}. Our experiments are all conducted on two Nvidia GeForce RTX 2080Ti GPUs.

4.3. Evaluation

We conduct the quantitative experiments on V-COCO, HICO-DET and HOI-A to demonstrate the effectiveness of our method. The results are shown in Tab. 1, Tab. 2 and Tab. 3 in comparison with other classical methods.

For V-COCO, we report our result in terms of AP_{agent}, AP^{S1}_{role} and AP^{S2}_{role}. Our method outperforms state-of-the-art method by 3.04% and 1.45% on AP_{agent} and AP^{S2}_{role} respectively. This shows that our method could not only achieve excellent HOI detection performance, but also promote human action analysis.

For HICO-DET, we report result on Default and Known Object settings. And for each setting, the results on Full, Rare and Non-Rare subset are shown in Tab. 2. Although some methods [12,31,35,37,41,42,45,46] adopt exter-nal human pose or linguistic knowledge or both to help HOI detection, our method outperforms them by a large margin especially on Full and Non-Rare set. In addition, compared to recently proposed global-context-aware methods [13,15–17,49], our method outperforms them on all subsets.

However, we only achieve competitive performance on AP^{S1}_{role} of V-COCO and Rare set of HICO-DET compared to IDN. Considering V-COCO is a smaller dataset than HICO-DET and Rare set of HICO-DET only has a few samples for each category, it means that our method needs relatively more samples to train a powerful model.

To show the adaptability of different datasets, we also test our method on HOI-A dataset. As is shown in Tab. 3, we achieve 3.33 mAP improvement compared to state-of-the-art PPDM. The result shows that our method can be applied to a variety of scenarios with Human-Object Interactions.

Following InteractNet [8], we also report Role AP for each interaction category and Average Role AP of all cat-egories on V-COCO [27]. The result is shown in Tab. 4. Our result outperforms result reported in [8] by a large margin. Besides, through column 2–3 and column 4–5 we find that our method gets 8.14% and 7.09% Average Role AP improvement on scenario_1 and scenario_2 when using ground truth (GT) human/object boxes instead of human /object boxes detected by Faster R-CNN. Thus, HOI detection with our framework could be further promoted once a better detector is adopted.

4.4. Ablation study

RGBM Generator. In Table 5a, we conduct an ablation study to show the effectiveness of RGBM generator. “RGB” represents not indicating actor position at all. “RGB + 255” represents indicating actor position by changing image structure of “RGB” channels. “RGBM” represents indicating actor position in the additional channel as proposed in the

Table 1
Result on V-COCO. Character * indicates that external knowledge is used.

Methods	AP^{agent}	$AP^{S_1}_{role}$	$AP^{S_2}_{role}$
InteractNet [8]	69.2	40.0	–
GPNN [29]	–	44.5	42.8
iCAN [13]	–	45.3	52.4
RPNN [30]	–	47.5	–
VCL [9]	–	48.3	–
TIN * [31]	–	48.7	54.2
Zhou et al. [32]	–	48.9	–
TIK [33]	–	48.7	–
PastaNet * [34]	–	51.0	57.5
DRG * [35]	–	21.0	–
Wei et al. [36]	70.3	42.0	–
ACP * [37]	–	53.2	–
IDN [38]	–	**53.3**	60.3
Ours	**73.47**	51.67	**61.75**

Table 2
Results on HICO-DET. For external knowledge, “P” indicates human pose and “L” indicates linguistic knowledge.

Methods	External knowledge	Default			Known Object		
		Full	Rare	Non-Rare	Full	Rare	Non-Rare
iCAN [13]		14.84	10.45	16.15	16.26	11.33	17.73
Wang et al. [16]		16.24	11.16	17.75	17.73	12.78	19.21
PMFNet [41]	P	17.46	15.65	18.00	20.34	17.47	21.20
No-Frills [42]	L	17.18	12.17	18.68	–	–	–
TIN [31]	P	17.22	13.51	18.32	19.38	15.38	20.57
CHG [43]		17.57	16.85	17.78	21.00	20.74	21.08
UnionDet [44]		17.58	11.52	19.33	19.76	14.68	21.27
Peyre et al. [45]	L	19.40	14.63	20.87	–	–	–
VSGNet [14]		19.80	16.05	20.91	–	–	–
FCMNet [46]	P + L	20.41	17.34	21.56	22.04	18.97	23.12
ACP [37]	P + L	20.59	15.92	21.98	–	–	–
Bansal et al. [47]	L	21.96	16.43	23.62	–	–	–
DRG [35]	L	24.53	19.47	26.04	27.98	23.11	29.43
IDN [38]		26.29	**22.61**	27.39	28.24	**24.47**	29.37
ConsNet-F [12]	L	24.39	17.10	26.56	–	–	–
IPNet [17]		19.56	12.79	21.58	22.05	15.77	23.92
PPDM [15]		21.73	13.78	24.10	24.58	16.65	26.84
HOTR [48]		25.10	17.34	27.42	–	–	–
Zou et al. [49]		26.61	19.15	28.84	29.13	20.98	31.57
Ours		**27.39**	21.34	**29.20**	**30.87**	24.20	**32.87**

Table 3
Result on HOI-A test set.

Methods	mAP (%)
C-HOI [32]	66.04
iCAN [13]	44.23
TIN [31]	48.64
PPDM [15]	71.23
Ours	**74.56**

Table 4
Results of each verb category on V-COCO test set. We report the result on scenario_1, and scenario_2. "Detector" means using detected human/object boxes. GT means using ground truth human/object boxes.

Action	scenario_1		scenario_2	
	Detector	GT	Detector	GT
hold-obj	36.52	41.83	55.20	62.21
sit-instr	29.25	40.25	55.58	63.93
ride-instr	71.86	79.41	77.68	83.42
look-obj	38.82	40.23	51.59	53.64
hit-instr	77.11	83.80	83.37	91.07
hit-obj	49.24	55.27	56.31	60.49
eat-obj	43.80	50.49	70.35	70.26
eat-instr	6.68	13.22	29.31	32.50
jump-instr	56.97	60.85	60.08	61.33
lay-instr	32.44	50.59	40.07	55.64
talk on phone-instr	58.77	78.20	66.05	90.66
carry-obj	39.24	41.72	43.58	45.75
throw-obj	49.60	51.13	53.82	55.60
catch-obj	46.74	46.22	57.19	57.21
cut-instr	47.70	66.52	55.39	75.36
cut-obj	39.96	43.83	54.21	59.43
work on computer-instr	63.75	73.79	70.51	79.67
ski-instr	52.76	68.56	65.57	80.27
surf-instr	81.31	89.90	84.90	93.44
skateboard-instr	89.27	93.65	92.89	96.71
drink-instr	34.91	43.19	37.10	43.19
kick-obj	75.98	76.14	85.39	85.30
read-obj	37.64	57.31	49.69	62.10
snowboard-instr	79.84	89.32	86.11	93.02
Average Role AP	51.67	59.81	61.75	68.84

section of METHODOLOGY. We see a successive improvement from Line 1 to Line 3. This proves that "RGBM" is an effective way to guide interaction reasoning.

Actor Branch. To validate the necessity of our Actor Branch, we conduct an ablation study shown in Table 5b. We start from a baseline without actor branch in the training.

and testing process, and obtains 25.60 mAP in full subsets. While we add the actor branch on the training procedure and testing procedure, the performance obtains an improvement with 0.84 and 1.79 respectively. This proves that Actor Branch can guide the Object Branch to learn better. And directly fusing prediction of Actor Branch and Object Branch can further improve performance.

Additional Category. We further evaluate the effective-ness of the additional category, *i.e.*, w/o-interaction cate-gory. This category is considered to add an explicit back-ground category to split the interaction verb scores and non-interaction verb scores (though Non-interaction cate-gory is included in HICO-DET, it means that no interaction appears on the image). We add it to Actor Branch and Object Branch successively and obtains a slight improve-ment of 0.39 and 0.47. But when both branches are added simultaneously, the performance achieves a bit lower due to the difficult joint optimization of both branches.

Weighted Cross Entropy Loss. The proposed weighted cross entropy loss consists of two types of weights, *i.e.*, Hanning Weight, and Scale Weight. To validate their effectiveness, we conduct an ablation study as shown in Table 5d. We can see both loss weights could help to improve performance and the performance could be further improved when they are used together.

Backbone. To further verify the robustness of the pro-posed framework, we conduct experiments as shown in Table 5e to compare with existing CNN-Based methods when using different backbones. We compare with VCL [9], PastaNet [34] and PPDM. VCL and PastaNet are both crop-based methods. Besides, PastaNet uses human pose and linguistic knowledge to guide interaction prediction. PPDM is one of the first methods that use non-crop features to predict interactions.

Results in line 1–3 and line 4–6 show that our method brings significant improvements when using the same backbone (no matter Res50 or DLA34), except for on Rare set compared to PastaNet (which is with the help of the external knowledge for improving the performance on Rare set). We further achieve state-of-the-art results when using HRNet-W32 (line 5). The results show the adapt-ability of our framework to different backbones.

Table 5
Ablation studies on HICO-DET.

(a) Effectiveness of RGBM Generator. We try different inputs of verb prediction network (Fig. 2b). "RGB" means using the original image as input. "RGB + 255" means add 255 to the intensity of pixels that lie inside actor bounding box. "RGBM" means using proposed RGBM data as input.

Input	Full	Rare	Non-Rare
RGB	21.88	16.84	23.39
RGB + 255	24.51	18.77	26.23
RGBM	26.18	19.88	29.06

(b) Impact of Actor Branch. The Baseline only uses Object Branch for training and testing. Then we add Actor Branch during training procedure and further fuse prediction of Actor Branch and Object Branch when testing. The results indicate that just learning with Actor Branch could promote the performance. Fusing prediction of both branches could bring further improvement

Baseline	Actor-Branch		Full	Rare	Non-Rare
	Train	Test			
✓	✓		25.60	18.48	27.73
✓			26.44	18.97	28.68
✓	✓	✓	27.39	21.34	29.20

(c) Effectiveness of Additional Category. Baseline uses the original verb categories as supervision for both branches. A^+ and O^+ means adding w/o-interaction category as supervision to Actor Branch and Object Branch respectively. We achieve best improvement when just adding w/o-interaction category to Object Branch

Baseline	$_A+$	$_O+$	Full	Rare	Non-Rare
✓			26.92	21.18	28.64
✓	✓		27.33	21.33	29.13
✓		✓	27.39	21.34	29.20
✓	✓	✓	27.07	20.42	29.07

(d) Effectiveness of Weighted Cross Entropy Loss. Standard binary cross entropy loss is used by default. We show the results when using Hanning Weight and Scale Weight alone as well as using both together. It's obvious that both loss weights could help to improve performance and the performance could be further improved when they are used together

Hanning Weight	Scale Weight	Full	Rare	Non-Rare
		25.30	18.41	27.36
✓		25.61	19.36	27.53
	✓	25.45	19.10	27.41
✓	✓	25.83	19.79	27.60

(e) Results on different backbones. Our method outperforms existing methods when using the same backbone. We achieves best results when using HRNet-W32

Methods	Backbone	Full	Rare	Non-Rare
VCL [9]	Res-50	23.63	17.21	25.55
PastaNet [34]	Res-50	22.65	**21.17**	23.09
Ours	Res-50	**24.58**	17.71	**26.63**
PPDM [15]	DLA34	20.29	13.06	22.45
Ours	DLA34	**26.18**	**19.88**	**29.06**
Ours	HRNet-W32	**27.39**	**21.34**	**29.20**

4.5. Analysis and discussion

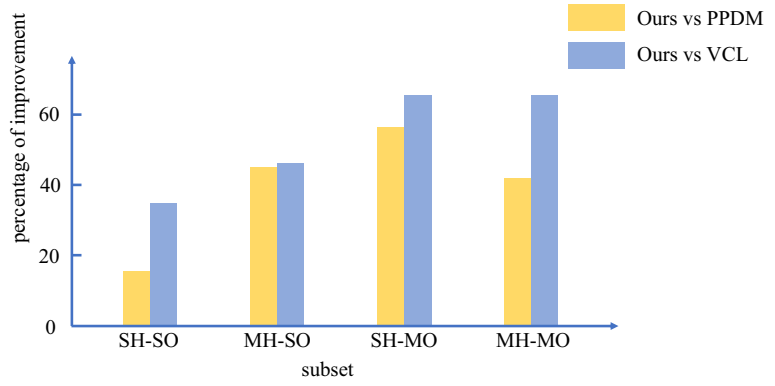
Results on simple and complex subsets. To evaluate the robustness of our proposed idea for data with different complexity, we split the HICO-DET into four subsets according to the number of annotated persons and ob-jects, *i.e.*, multi-human & multi-object (MH-MO), single-human & multi-object (SH-MO), multi-human & single-object (MH-SO) and single-human & single-object (SH-SO). We compare our proposed framework with PPDM and VCL on these four subsets in Table 6. Results intuitively show that our method consistently obtains

Table 6
Results on different subsets. These subsets are generated according to the number of annotated persons and objects in the image, *i.e.*, single-human & single-object (SH-SO), multi-human & single-object (MH-SO), single-human & multi-object (SH-MO) and multi-human & multi-object (MH-MO). Our method obtains consistent improvement on all subsets and is especially more robust in complex scenes.

(a) The results (mAP) on subsets of HICO-DET. (*%↑) denotes the increasing percentage we achieve compared to PPDM and VCL

	SH-SO	MH-SO	SH-MO	MH-MO
PPDM	29.03(**15%**↑)	17.17(**44%**↑)	15.94(**55%**↑)	13.06(**41%**↑)
VCL	24.78(**34%**↑)	17.14(**45%**↑)	15.04(**64%**↑)	11.23(**64%**↑)
OURS	33.27	24.80	24.70	18.47

(b) The visualization version of results reported in the above Table. From this figure we can intuitively observe that improvements achieved by our framework of complex subsets (MH-SO, SH-MO, MH-MO) are higher than the simple one (SH-SO)



better performance on all four subsets. Especially for complex data (SH-MO, MH-SO, MH-MO), our method achieves more improve ment. That is, our method is consistent effective and more suitable for complex scenes.

Qualitative Result**.** Some qualitative results compared with PPDM and VCL in shown in Fig. 5. We select several images in the HICO-DET and V-COCO datasets to show the ability of our proposed framework of addressing the multiple interaction occurrence issues. These cases are quite complex and easy to cause ambiguities for crop-based method, and highly-coupled approach. As illustrated in Fig. 5, our approach handles the complex crowded scenes, which are failed by the two other approaches. And the Object Scoremap shown in the last row further validates the effectiveness of our model to capture the interactions between actor and objects.

Moreover, we also illustrate the heatmap results of the Object Branch for different actors in the same image as shown in Fig. 6. The

Fig. 5. Visualization comparison results of our method and PPDM [15] and VCL [9] on some complex scenes from HICO-DET and V-COCO datasets. From top to bottom: The results of PPDM and VCL, our detection result, ground truth, and object scoremap predicted by our Object Branch. Clearly, in each specific HOI detection, high scores would be regressed for the human and its interacted objects. Thus, our approach obtains better performance on these complex cases, which validates the effectiveness of our framework to handle ambiguity issues.

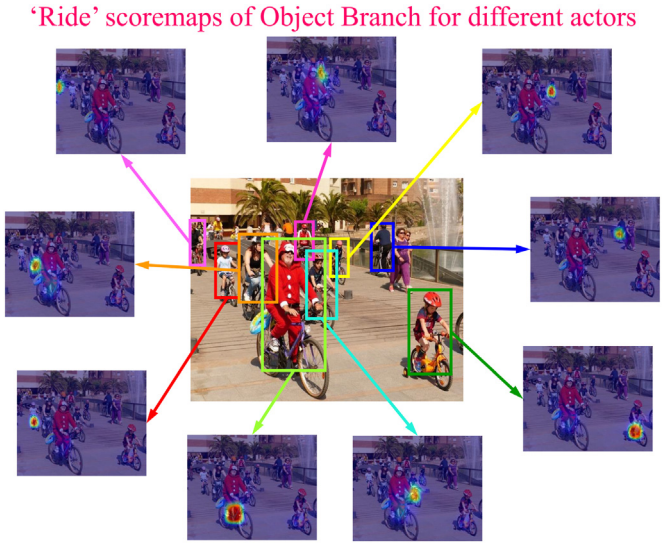


Fig. 6. The heatmap results of Object Branch for different actors in the same image. Boxes with different colors mean different detected persons. When different individuals serve as the actor, the resulting heatmaps are shown where the colored arrows point. It has been observed that our model has the ability to localize the corresponding interacting regions for each actor accurately in complex scenes with multiple interactions.

high consistency between actors and object regions demonstrates that our approach has a powerful ability to disambiguate multiple complex interactions in the same image.

5. Conclusion

In this paper, we have developed an effective actor-centric approach for human-object interaction detection. Our approach formulates the task as a pixel-wise prediction problem, and learns non-local features by introducing binary masks of persons. The relationship between one person and multiple objects of the entire image has been explored to promote the interaction reasoning in a unified framework. For enhancing the learning effect under the proposed framework, we further introduce a Weighted Cross Entropy Loss consisting of Hanning weight and Scale weight. Experimental results have shown that our approach achieves state-of-the-art performance in both HICO-DET and V-COCO datasets, especially on complex images with multiple persons/objects.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work is supported by the National Natural Science Foundation of China (NSFC) grant 62176100, the Central Guidance on Local Science and Technology Development Fund of Hubei Province grant 2021BEE056.

References

[1] C.-Y. Chen, K. Grauman, Predicting the location of "inter-actees" in novel human-object interactions, Asian Conference on Computer Vision, Springer 2014, pp. 351–367.
[2] B. Yao, L. Fei-Fei, Modeling mutual context of object and human pose in human-object interaction activities, 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, IEEE 2010, pp. 17–24.
[3] C. Gao, S. Liu, D. Zhu, Q. Liu, J. Cao, H. He, R. He, S. Yan, Interactgan: learning to generate human-object interaction, Proceedings of the 28th ACM International Conference on Mul timedia 2020, pp. 165–173.
[4] E.E. Aksoy, A. Abramov, J. Dorr, K. Ning, B. Dellen, F. Worgotter, Learning the semantics of object–action relations by observation, Int. J. Robot. Res. 30 (10) (2011) 1229–1249.
[5] F. Worgotter, E.E. Aksoy, N. Krüger, J. Piater, A. Ude, M. Tamosiunaite, A simple ontology of manipulation actions based on hand-object relations, IEEE Trans. Auton. Ment. Dev. 5 (2) (2013) 117–134.
[6] B.D. Argall, S. Chernova, M. Veloso, B. Browning, A survey of robot learning from demonstration, Robot. Auton. Syst. 57 (5) (2009) 469–483.
[7] Y. Yang, C. Fermuller, Y. Aloimonos, Detection of manipu-lation action consequences (mac), Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2013, pp. 2563–2570.
[8] G. Gkioxari, R. Girshick, P. Dollar, K. He, Detecting and recognizing human-object interactions, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2018, pp. 8359–8367.
[9] Z. Hou, X. Peng, Y. Qiao, D. Tao, Visual compositional learning for human-object interaction detection, European Conference on Computer Vision, Springer 2020, pp. 584–600.
[10] Y.-W. Chao, Y. Liu, X. Liu, H. Zeng, J. Deng, Learning to detect human-object interactions, 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), IEEE 2018, pp. 381–389.
[11] B. Xu, Y. Wong, J. Li, Q. Zhao, M.S. Kankanhalli, Learning to detect human-object interactions with knowledge, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019.
[12] Y. Liu, J. Yuan, C.W. Chen, Consnet: learning consistency graph for zero-shot human-object interaction detection, Proceedings of the 28th ACM International Conference on Multi-media 2020, pp. 4235–4243.
[13] C. Gao, Y. Zou, J.-B. Huang, ican: Instance-Centric Attention Network for Human-Object Interaction Detection, 2022 arXiv preprint arXiv:1808.10437.
[14] O. Ulutan, A. Iftekhar, B.S. Manjunath, Vsgnet: spatial at-tention network for detecting human object interactions using graph convolutions, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 2020, pp. 13617–13626.
[15] Y. Liao, S. Liu, F. Wang, Y. Chen, C. Qian, J. Feng, Ppdm: parallel point detection and matching for real-time human-object interaction detection, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 2020, pp. 482–490.
[16] T. Wang, R.M. Anwer, M.H. Khan, F.S. Khan, Y. Pang, L. Shao, J. Laaksonen, Deep contextual attention for human-object interaction detection, Proceedings of the IEEE/CVF International Conference on Computer Vision 2019, pp. 5694–5702.
[17] T. Wang, T. Yang, M. Danelljan, F.S. Khan, X. Zhang, J. Sun, Learning human-object interaction detection using interaction points, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 2020, pp. 4116–4125.
[18] X. Zhong, X. Qu, C. Ding, D. Tao, Glance and gaze: Inferring action-aware points for one-stage human-object interaction detection, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 2021, pp. 13234–13243.
[19] A. Zhang, Y. Liao, S. Liu, M. Lu, Y. Wang, C. Gao, X. Li, Mining the benefits of two-stage and one-stage hoi detection, Thirty-Fifth Conference on Neural Information Processing Systems, 2021.
[20] M. Chen, Y. Liao, S. Liu, Z. Chen, F. Wang, C. Qian, Reformulating hoi detection as adaptive set prediction, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 2021, pp. 9004–9013.
[21] Z. Li, C. Zou, Y. Zhao, B. Li, S. Zhong, Improving Human-Object Interaction Detection via Phrase Learning and Label Composition, 2022 arXiv preprint arXiv:2112.07383.
[22] A. Newell, K. Yang, J. Deng, Stacked hourglass networks for human pose estimation, European Conference on Computer Vision, Springer 2016, pp. 483–499.
[23] H. Law, J. Deng, Cornernet: detecting objects as paired keypoints, Proceedings of the European Conference on Computer Vision (ECCV) 2018, pp. 734–750.
[24] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, S. Belongie, Feature pyramid networks for object detection, Proceedings of the IEEE Conference on Computer Vision and pattern recognition 2017, pp. 2117–2125.
[25] F. Yu, D. Wang, E. Shelhamer, T. Darrell, Deep layer aggregation, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2018, pp. 2403–2412.
[26] K. Sun, B. Xiao, D. Liu, J. Wang, Deep high-resolution representation learning for human pose estimation, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 2019, pp. 5693–5703.
[27] S. Gupta, J. Malik, Visual Semantic Role Labeling, 2022 arXiv preprint arXiv:1505.04474.
[28] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ra-manan, P. Dollar, C.L. Zitnick, Microsoft coco: common objects in context, European Conference on Computer Vision, Springer 2014, pp. 740–755.
[29] S. Qi, W. Wang, B. Jia, J. Shen, S.-C. Zhu, Learning human-object interactions by graph parsing neural networks, Proceedings of the European Conference on Computer Vision (ECCV) 2018, pp. 401–417.
[30] P. Zhou, M. Chi, Relation parsing neural network for human-object interaction detection, Proceedings of the IEEE/CVF International Conference on Computer Vision 2019, pp. 843–851.
[31] Y.-L. Li, S. Zhou, X. Huang, L. Xu, Z. Ma, H.-S. Fang, Y. Wang, C. Lu, Transferable interactiveness knowledge for human-object interaction detection, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2019, pp. 3585–3594.
[32] T. Zhou, W. Wang, S. Qi, H. Ling, J. Shen, Cascaded human-object interaction recognition, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recog nition 2020, pp. 4263–4272.
[33] Y.-L. Li, S. Zhou, X. Huang, L. Xu, Z. Ma, H.-S. Fang, Y. Wang, C. Lu, Transferable interactiveness knowledge for human-object interaction detection, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 2019, pp. 3585–3594.

[34] Y.-L. Li, L. Xu, X. Liu, X. Huang, Y. Xu, S. Wang, H.-S. Fang, Z. Ma, M. Chen, C. Lu, Pastanet: toward human activity knowledge engine, Proceedings of the IEEE/CVF Con-ference on Computer Vision and Pattern Recognition 2020, pp. 382–391.
[35] C. Gao, J. Xu, Y. Zou, J.-B. Huang, Drg: dual relation graph for human-object interaction detection, European Conference on Computer Vision, Springer 2020, pp. 696–712.
[36] W. Feng, W. Liu, T. Li, J. Peng, C. Qian, X. Hu, Turbo learn-ing framework for human-object interactions recognition and human pose estimation, Proceedings of the AAAI Conference on Artificial Intelligence 2019, pp. 898–905.
[37] D.-J. Kim, X. Sun, J. Choi, S. Lin, I.S. Kweon, Detecting Human-Object Interactions with Action Co-Occurrence Priors, 2022 arXiv preprint arXiv:2007.08728.
[38] Y.-L. Li, X. Liu, X. Wu, Y. Li, C. Lu, Hoi analysis: integrating and decomposing human-object interaction, Advances in Neural Information Processing Systems 2022, p. 33.
[39] S. Ren, K. He, R. Girshick, J. Sun, Faster r-cnn: towards real-time object detection with region proposal networks, IEEE Trans. Pattern Anal. Mach. Intell. 39 (6) (2016) 1137–1149.
[40] K. Chen, J. Wang, J. Pang, Y. Cao, Y. Xiong, X. Li, S. Sun, W. Feng, Z. Liu, J. Xu, et al., Mmdetection: Open mmlab Detection Toolbox and Benchmark, 2022 arXiv preprint arXiv:1906.07155.
[41] B. Wan, D. Zhou, Y. Liu, R. Li, X. He, Pose-aware multi-level feature network for human object interaction detection, Proceedings of the IEEE/CVF International Conference on Com puter Vision 2019, pp. 9469–9478.
[42] T. Gupta, A. Schwing, D. Hoiem, No-frills human-object in-teraction detection: factorization, layout encodings, and training techniques, Proceedings of the IEEE/CVF International Conference on Computer Vision 2019, pp. 9677–9685.
[43] H. Wang, W.-s. Zheng, L. Yingbiao, Contextual Heterogeneous Graph Network for Human-Object Interaction Detection, 2022 arXiv preprint arXiv:2010.10001.
[44] B. Kim, T. Choi, J. Kang, H.J. Kim, Uniondet: union-level detector towards real-time human-object interaction detection, European Conference on Computer Vision, Springer 2020, pp. 498–514.
[45] J. Peyre, I. Laptev, C. Schmid, J. Sivic, Detecting unseen visual relations using analogies, Proceedings of the IEEE International Conference on Computer Vision 2019, pp. 1981–1990.
[46] Y. Liu, Q. Chen, A. Zisserman, Amplifying key cues for human-object-interaction detection, European Conference on Computer Vision, Springer 2020, pp. 248–265.
[47] A. Bansal, S.S. Rambhatla, A. Shrivastava, R. Chellappa, Detecting human-object interactions via functional generalization, AAAI (2020) 10460–10469.
[48] B. Kim, J. Lee, J. Kang, E.-S. Kim, H.J. Kim, Hotr: end-to-end human-object interaction detection with transformers, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 2021, pp. 74–83.
[49] C. Zou, B. Wang, Y. Hu, J. Liu, Q. Wu, Y. Zhao, B. Li, C. Zhang, C. Zhang, Y. Wei, et al., End-to-end human object interaction detection with hoi transformer, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 2021, pp. 11825–11834.
[50] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al., Imagenet large scale visual recognition challenge, Int. J. Comput. Vis. 115 (3) (2015) 211–252.
[51] D.P. Kingma, J. Ba, Adam: A Method for Stochastic Optimization, 2022 arXiv preprint arXiv:1412.6980.