2023 IEEE International Conference on Robotics and Automation (ICRA 2023)
May 29 - June 2, 2023. London, UK

SGPT: The Secondary Path Guides the Primary Path in Transformers for HOI Detection

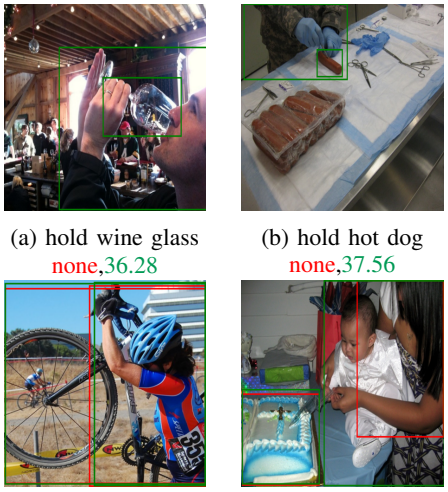
Sixian Chan[1,2], Weixiang Wang[1], Zhanpeng Shao[3] and Cong Bai[1]

***Abstract*— HOI detection is essential for human-computer interaction, especially in behavior detection and robot manipulation. Existing mainstream transformer methods of HOI detection are focused on single-stream detection only, e.g., $image \rightarrow HOI(\mathcal{P}_1)$, or $image \rightarrow HO \rightarrow I(\mathcal{P}_2)$. Both paths have their own characteristics of concern, so we propose a novel method, using the Secondary path ($\mathcal{P}_2$) Guides the Primary path ($\mathcal{P}_1$) in Transformers (SGPT). SGPT contains two core modules: the Dual-Path Consistency (DPC) module and the Instance Interaction Attention (IIA) module. DPC keeps human, object and interaction consistent on the dual-path and lets $\mathcal{P}_2$ guide $\mathcal{P}_1$ to learn more meaningful features. IIA fuses human and object to enhance interaction in $\mathcal{P}_2$, which allows instance to constrain interaction. Our proposed dual-path are employed during training, and only the $\mathcal{P}_1$ path is used for inference. Hence, SGPT improves generalization without increasing model capacity in HICO-DET and V-COCO datasets compared to the state-of-the-arts. The code of this work is available at https://github.com/visualVk/sgpt.git.**

I. Introduction

Human-Object Interaction (HOI) detection plays a significant role in understanding high-level information. Therefore, the improvement of HOI detection has significance for human-computer interaction. In robot manipulation [1, 2], the robot needs to imitate humans' actions and complete operations, or infer the later actions from our actions and complete such operations. In behavior monitoring [3], robots need to be able to detect dangerous behaviors and warn or remedy humans directly. Then, some wrong cooperative behaviors can be avoided by monitoring and understanding human interactions for robots. HOI detection requires discovering as well as possible the correct HOI triplets: $\langle human, object, interaction \rangle$. It includes locations of humans, categories of objects and interactions happened on corresponding pairs of humans and objects, respectively.

Overall, detecting HOI triplets can be divided into locating and classifying for human and object, pairing them correctly and reasoning right interaction on pairs of human and object. So we can simplify this task into two parts: 1) location of human and object and recognition categories of object same as previous detection task; 2) associating human with object corresponding as HO pairs and discriminating interactions of pairs. We revisit approaches, two-stage and one-stage, how to finish two sections on HOI detection.



(a) hold wine glass
none,36.28

(b) hold hot dog
none,37.56

(c) inspect bicycle
33.40, 49.10

(d) cut cake
23.27, 56.19

Fig. 1: Results of qualitative analysis compared to QPIC [4]. Each image contains two colors of boxes and scores of interaction. The green is our network prediction and the red is the predicted by QPIC. None means that the network fails to detect the correct box and the interaction.

For two-stage, this framework [5]–[9] uses a great enough detector [10] and freezes its parameters to find all humans and objects in the image. Then, humans and objects are paired by enumerating, such as [5, 6]. However, these simple HO pairs do not have strong relationships. So literature [7]–[9] introduces GCN [11] to understand relations. Humans or objects ignored by the detector never show in the ground truth, although GCN strengthens relationships of correct pairs. In addition, this approach pays too much computation on the process of matching.

For one-stage, this framework [4, 12]–[15] associates humans with objects early (before locating and classifying) and parallelly classifies and locates of humans, objects and interactions. The conventional one-stage [13, 16] costs much computation on associating human and object with interaction. But some approaches work out this problem, they [4, 14, 15] make one pair into one token, which can be applied for classification and regression of human, object and interaction. However, some methods based on transformer [17]–[19] will obtain some useless HOI triplets such as "human ride headset".

[1]Sixian Chan, Weixiang Wang and Cong Bai are with College of Computer Science and Technology, Zhejiang University of Technology, Zhejiang 310023, China sxchan@zjut.edu.cn, feifeiyue@gmail.com and Cong Bai is corresponding author:congbai@zjut.edu.cn

[2] Hubei Key Laboratory of Intelligent Vision Based Monitoring for Hydroelectric Engineering; the College of Computer and Information at China Three Gorges University, Yichang 443002, China

[3]Zhanpeng Shao is with College of Information Science and Engineering, Hunan Normal University, Changsha 410081, China zpshao@hunnu.edu.cn

2023 IEEE International Conference on Robotics and Automation (ICRA) | 979-8-3503-2365-8/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICRA48891.2023.10160329

Follow previous works [4, 9, 13]–[15, 20], we research detection path such as methods [4, 13]–[15] based on the primary path ($\mathcal{P}_1$) and methods [8, 20] based on secondary path $\mathcal{P}_2$. However, $\mathcal{P}_2$ belongs to the two-stage and can not be suited to one-stage. So a very simple but analogous two-stage cascade decoder is designed in our work. In order to make full use of the two methods, we propose Dual-Path Consistency (DPC) to associate $\mathcal{P}_1$ with $\mathcal{P}_2$. $\mathcal{P}_2$ guides $\mathcal{P}_1$ to focus on useful features. Nevertheless, we consider that this simple cascade decoder does not take full advantage of the human and object constraints interactions. Under this consideration, the Instance Interaction Attention (IIA) is proposed to overcome the above problem and enhance interaction features in $\mathcal{P}_2$. Although we propose dual-path, only the results of $\mathcal{P}_1$ are used as predictions. $\mathcal{P}_2$ is only as an auxiliary to $\mathcal{P}_1$. In inference, $\mathcal{P}_2$ will be removed, so our model can guarantee generalization and also take into account the inference efficiency.

In this work, our contributions can be summarized in the following:

1) We propose the Dual-Path Consistency (DPC) to take full advantage of the two detection paths for HOI Detection, keeping human, object and interaction consistent on the dual-path and lets $\mathcal{P}_2$ guide $\mathcal{P}_1$ to learn more meaningful features.
2) We design the Instance Interaction Attention (IIA) to improve the two-stage cascade decoder of $\mathcal{P}_2$, which fuses humans and objects into interactions to strengthen the instance for interaction restrictions.
3) Extensive experiments are conducted on both V-COCO and HICO-DET datasets. The experimental results show that our method achieves state-of-the-art performance.

II. Related Work

A. Two-stream Architecture

Two-stream networks are decomposed into spatial and temporal streams. Each stream extract features separately for late fusing in the field of video recognition [21]. After two streams are used, it is clear that each stream focuses on only one thing. It is very obvious that: 1) the features of two streams are particularly characteristic; and 2) the two features that can be fused by simple fusion operations (weighting [22], MLP [23]) will yield more informative features. Therefore, we follow the two-stream network architecture and propose a Dual-Path Consistency learning strategy with each path corresponding to a specific detector.

B. HOI Detection

HOI Detection is mainly grouped into two-stage and one-stage approaches. The two-stage approaches [4, 9, 13, 15, 20] rely on an advanced detector, like Faster-RCNN [10] or DETR [18]. Once objects are detected perfectly in the current image, the HO pairs can be built. Finally, an interaction detection is performed on the HO pairs. The general pairing will be done with the help of GCN [11]. But it appears to have the following disadvantages: 1) there are many forced associations of pairings that are meaningless; 2) when there are many objects in the image, it may combine approximate N^2 pairs of HO, which will consume a lot of computational resources. One-stage methods [4, 12]–[15] are able to form a pairing of targets without matching. These pairs are more meaningful than the two-stage's. Since the one-stage approach is now almost always based on the transformer, the number of pairs is a constant K, which alleviates the problem of computational resource consumption of the two-stage and removes the post-process of matching HO pairs and interactions in conventional one-stage [16].

C. Consistency Learning

Consistency learning is widely used in the computer vision [24]–[27], especially in multi-modal [24] and semi-supervised [25]. In multi-modal, consistent learning can learn common features of multiple modalities, such as instance alignment in the ReID [24]. In semi-supervised, it is often used to maintain consistency with the intact features after dropout features randomly or to mask the image to maintain feature consistency. This allows the final extracted features to have better noise immunity.

III. Method

A. Overall architecture

In Fig. 2, our overall architecture includes a Visual Encoder and HOI decoder. We propose DPC module and IIA module mainly in the HOI decoder to improve the generalization.

Visual Encoder. The design of our visual encoder is inspired by QPIC [4]. In detail, the input is a image $I \in R^{H\times W\times C}$. The feature map $I' \in R^{H'\times W'\times C'}$ is obtained by CNN [28] encoder, which is flattened and then projected to a lower dimensional of D to obtain $X \in R^{H'W'\times D}$. This lower dimensional feature is fed into the transformer encoder with sinusoidal positional embedding to obtain the visual feature $V_e \in R^{H'W'\times D}$.

HOI Decoder. We share the same decoder [29, 30] for our dual-path. We employ a 3-layer MLP for the classification and a fully-connected layer for the location. $\mathcal{P}_1$ and $\mathcal{P}_2$ use their learnable query embedding $Q_{hoi} \in R^{N\times D}$ and $Q_{ins} \in R^{2N\times D}$, respectively. $\mathcal{P}_1$ is to pass a decoder to get the learned HOI features $e^{\mathcal{P}_1}_{hoi} \in R^{N\times D}$. $\mathcal{P}_2$ first passes a decoder to get HO features $e^{\mathcal{P}_2}_{ins} \in R^{N\times D}$, then passes another decoder to get I features $e^{\mathcal{P}_2}_{inter} \in R^{N\times D}$, and finally gets the prediction result $(b_h, b_o, c_o, c_{inter}) \in Pr$.

TABLE I: Detection paths

$\mathcal{P}_1$:	$x \to HOI$
$\mathcal{P}_2$:	$x \to HO \to I$
$\mathcal{P}_3$:	$x \to HI \to O$
$\mathcal{P}_4$:	$x \to OI \to H$
$\mathcal{P}_5$:	$x \to I \to HO$
$\mathcal{P}_6$:	$x \to H \to OI$
$\mathcal{P}_7$:	$x \to O \to HI$

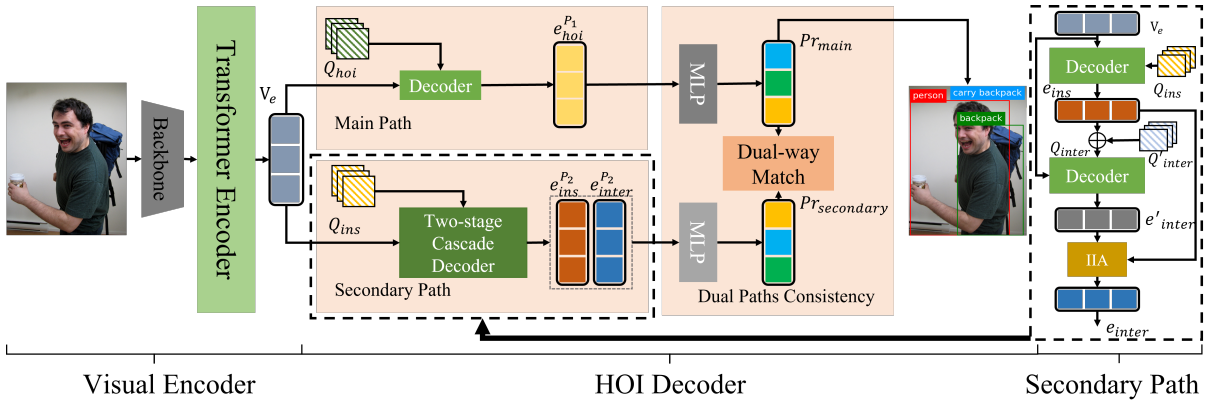


Fig. 2: Overall network architecture. The image is encoded to get V_e. V_e is fed into dual-path to receive tokens respectively. After dual-way matching for tokens, the network outputs the predictions$\langle human, object, interaction\rangle$.

B. Dual-Path Consistency

Path selection. We find that whether it is a CNN-based [8, 13, 20] or transformer-based [4, 14, 15] HOI detection, both of them have multiple path detection interactions, such as Table I.

$\mathcal{P}_1$ and $\mathcal{P}_2$ are used for our Dual-Path Consistency. We do not use the remaining five paths because the interaction is affected by the subtle movement of humans and objects, as shown in Fig. 3. However, human and object have strict bounding-box and classification labels, which would be precise in supervised training. So using interaction to infer human or object will result in a relatively large error. In addition, we assume that $Set(\mathcal{P}_i)$ is a positive and negative sample volume in $\mathcal{P}_i$. The above seven paths are related as follows:

$$\begin{aligned} Set(\mathcal{P}_1) &= Set(\mathcal{P}_2) + Set(\mathcal{P}_3) + Set(\mathcal{P}_4) + Set(\mathcal{P}_5) \\ &\quad + Set(\mathcal{P}_6) + Set(\mathcal{P}_7) \end{aligned} \tag{1}$$

In Equation 1, maintaining consistency is actually using the $\mathcal{P}_2$-$\mathcal{P}_7$ constraint $\mathcal{P}_1$. This constraint will not only ensure that the positive samples are consistent, but also make the negative samples consistent. At the same time, since the number of queries is a constant K, it means that there is no way for predicting the full set of interactions. But $\mathcal{P}_3$-$\mathcal{P}_7$ with too much noise will predict more negative samples. If all paths are used, it will lead to too much focus on negative samples. Moreover, the positive samples are only expected to remain consistent. Hence, we choose paths $\mathcal{P}_1$ and $\mathcal{P}_2$ to establish our dual-path.

Dual-path design. After the paths are selected, it is important that how to design the dual-path to achieve the detection results with different yet having a more reasonable overlap from each path. To make the results of the paths different, different query embedding structures are designed. For $\mathcal{P}_1$, we use the same query embedding $Q_{hoi} \in R^{N\times D}$ for human, object and interaction (in Fig. 4a). For $\mathcal{P}_2$, human and object each use a query embedding $Q_h, Q_o \in R^{N\times D}$ as shown in Fig. 4b. Then, we generate instance query embedding $Q_{ins} \in R^{2N\times D}$ by concatenating Q_h and Q_o. The interaction query

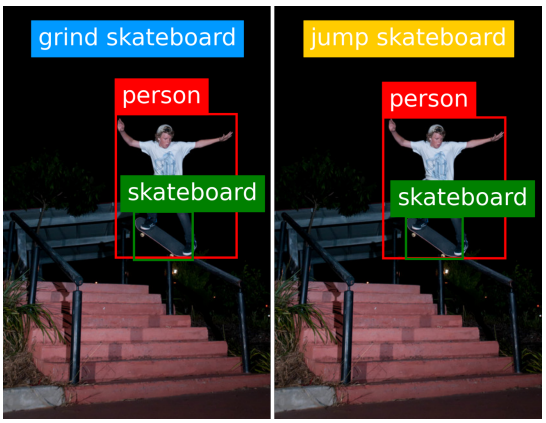


Fig. 3: The effect of subtle differences of object borders on interaction

embedding $Q_{inter} \in R^{N\times D}$ with position embedding $e_{ins} \in R^{N\times D}$ from decoder with Q_{ins}. The corresponding inference schemes of $\mathcal{P}_1$, $\mathcal{P}_2$ can be written in more formal terms for triplet result $\langle human, object, interaction\rangle$:

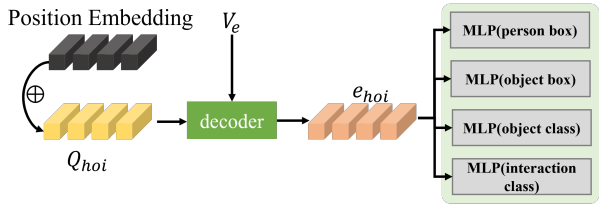
$$\begin{cases} y_h^{\mathcal{P}_1} = MLP_h^{\mathcal{P}_1}(f_{1,1}(x, V_e)) \\ y_o^{\mathcal{P}_1} = MLP_o^{\mathcal{P}_1}(f_{1,1}(x, V_e)) \\ y_{inter}^{\mathcal{P}_1} = MLP_h^{\mathcal{P}_1}(f_{1,1}(x, V_e)) \end{cases} \tag{2}$$

$$\begin{cases} y_h^{\mathcal{P}_2} = MLP_h^{\mathcal{P}_2}(f_{2,1}(Q_h, V_e)) \\ y_o^{\mathcal{P}_2} = MLP_o^{\mathcal{P}_2}(f_{2,1}(Q_o, V_e)) \\ y_{inter}^{\mathcal{P}_2} = MLP_h^{\mathcal{P}_2}(f_{2,2}(f_{2,1}(Q_{ins}, V_e), V_e)) \end{cases} \tag{3}$$

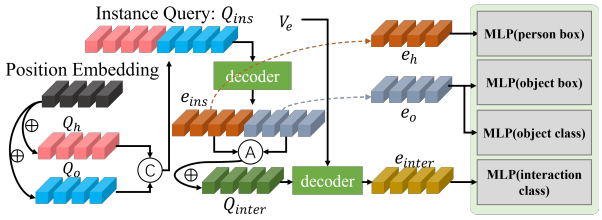
where, $y_i^{\mathcal{P}_j}$ is prediction result of i in j^{th} path, $i \in \{h, o, inter\}$ and $j \in (0, 1)$. $f_{i,j}$ denotes the j^{th} decoder in the i^{th} path. The visual feature V_e is obtained from the transformer encoder. $MLP(\cdot)$ is applied for classification or regression.

C. Instance Interaction Attention

We propose Instance Interaction Attention (IIA) to induce $\mathcal{P}_2$ learning with more prior meaningful interaction features. The simplest approach of the original HOI detection $\mathcal{P}_2$ is shown in Fig. 4b, which simply treats the information of human and object as position embedding to an interaction



(a) query embedding design of primary path



(b) query embedding design of secondary path

Fig. 4: Dual-path query embedding design. A means average operation.

query embedding $Q_{inter} \in R^{N\times D}$. It can learn the features of interaction by itself based on the known prior conditions of HO pairs. The way we fused is referred to the idea of channel attention [31] (in Fig. 6). We assume that HO pair features and interaction features are $e_{ins}, e_{inter} \in R^{N\times D}$, separately. Then, they will be concatenated together by computing a channel attention coefficient $\alpha \in R^{N\times D}$:

$$\alpha = sigmoid(MLP(concat(e_{ins}, e_{inter})) \tag{4}$$

where, the e_{ins} and e_{inter} are concatenated in the last dimension. $sigmoid(\cdot)$ is uesd to map the results to the range of 0 to 1. Then, we multiply the e_{inter} with α to obtain the channel-enhanced feature $e_{inter\_c} \in R^{N\times D}$. We let e_{ins} map to the interaction space to obtain $e'_{ins} \in R^{N\times D}$. Finally, we follow the skip connection of Resnet [28] and add e_{inter}, e_{inter_c} and e'_{ho} to obtain the instance-enhanced interaction feature $e \in R^{N\times D}$:

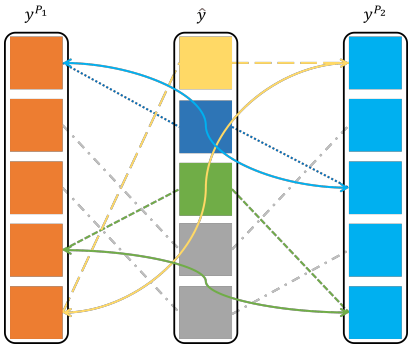


Fig. 5: Dual-path matching based on Ground Truth. $\hat{y}$ represents Ground Truth, where the gray blocks represent the filled blank blocks to align with the number of $\mathcal{P}_1$ and $\mathcal{P}_2$. The curved arrow solid line indicates the matching of the two prediction sets. The dashed line indicates the pairing of the prediction and $\hat{y}$, and different dashed lines indicate different category pairings.

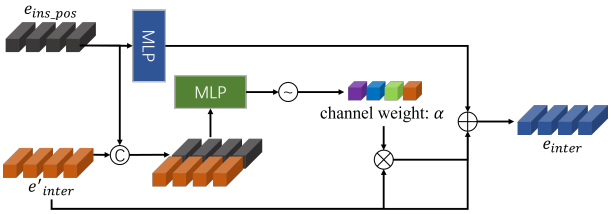


Fig. 6: The structure of Instance Interaction Attention. C, $\sim$, $\times$, + mean concatenation, regularization, dot product and sum, respectively.

$$e_{inter\_c} = \alpha \odot e_{inter} \tag{5}$$
$$e'_{ins} = MLP(e_{ins}) \tag{6}$$
$$e = e_{inter} + e_{inter\_c} + e'_{ins} \tag{7}$$

After our $\mathcal{P}_2$ has been enhanced by the features from IIA, the final prediction is computed as follows:

$$y_h^{\mathcal{P}_2} = MLP_h^{\mathcal{P}_2}(f_{2,1}(Q_h, V_e)) \tag{8}$$
$$y_o^{\mathcal{P}_2} = MLP_o^{\mathcal{P}_2}(f_{2,1}(Q_o, V_e)) \tag{9}$$
$$V_{e\_aug} = IIA(f_{2,2}(f_{2,1}(Q_{ins}, V_e), V_e)) \tag{10}$$
$$y_{inter}^{\mathcal{P}_2} = MLP_h^{\mathcal{P}_2}(V_{e\_aug}, f_{2,1}(Q_{ins}, V_e))) \tag{11}$$

where, to simplify the formula, $IIA(\cdot)$ is used instead of the above calculation.

D. Dual-way Constraints

Dual-way Matching. The prediction result is an unordered ensemble. If we directly calculate the loss by corresponding the two paths, it will lead to a terrible matching. However, we solve this problem by dual-way matching. The dual-path will each end up with a prediction set $y^i, i \in (\mathcal{P}_1, \mathcal{P}_2)$ and both will get the result $\sigma_{ij}(k), i \in index(y^k), j \in index(\hat{y}), k \in (\mathcal{P}_1, \mathcal{P}_2)$ that best matches the ground truth $\hat{y}$ by the Hungarian algorithm. Looking at the relationship of the pairing results in Fig. 5, the two paths can be correlated with the help of $\hat{y}$. For example, if we identify the pairing result of $\mathcal{P}_1$ as $\sigma_{1,2}(\mathcal{P}_1)$ and the pairing result of $\mathcal{P}_2$ as $\sigma_{3,2}(\mathcal{P}_2)$, we find that both of them pair up with 2 in $\hat{y}$. So we can assume that 1 in path $\mathcal{P}_1$ is paired with 3 in path $\mathcal{P}_2$. Following the above operation, we can convert $\sigma_{i,j}(\mathcal{P}_1)$ and $\sigma_{k,l}(\mathcal{P}_2)$ to $\delta_{i,k}(\mathcal{P}_1, \mathcal{P}_2)$ to represent the matching of the dual-path.

Consistency Loss. After we get the best matching relationship on the dual-path, we need to consider making its classification of object, interaction and the bounding-boxes of human and object consistent. Regarding the consistency of the classification, we use the KL divergence to measure the similarity between the dual-path to guarantee the consistency of the classification:

$$\mathcal{L}_{c_o} = \frac{1}{2\mathcal{N}} \sum_{j,k\in\delta} KLoss(z(c_o^{\mathcal{P}_1}, j) \,||\, z(c_o^{\mathcal{P}_2}, k)) \tag{12}$$
$$\mathcal{L}_{c_{inter}} = \frac{1}{2\mathcal{N}} \sum_{j,k\in\delta} KLoss(z(c_{inter}^{\mathcal{P}_1}, j) \,||\, z(c_{inter}^{\mathcal{P}_2}, k)) \tag{13}$$

TABLE II: Performance comparisons in HICO-DET and V-COCO. D, KO mean default and know object settings in HICO-DET. * denotes the results we reproduced using the official code. The highest results is **bolded**, while the second-best result is underlined.

Type	Method	HICO-DET						V-COCO	
		Full(D)	Rare(D)	Non-Rare(D)	Full(KO)	Rare(KO)	Non-Rare(KO)	Scenario 1	Scenario 2
Two Stage	Wang et al. [5]	17.57	16.85	17.78	21.00	20.74	21.08	52.70	-
	FCMNet [32]	20.40	17.30	21.60	22.04	18.97	23.12	53.10	-
	DRG [7]	24.53	19.47	26.04	27.98	23.11	29.43	51.00	-
	VCL [33]	19.43	16.55	20.29	-	-	-	48.30	-
	PD-Net [6]	22.37	17.61	23.79	26.86	21.70	28.44	52.60	-
	ACP [34]	20.59	15.92	21.98	-	-	-	52.98	-
	SABRA [35]	26.09	16.29	29.02	31.08	23.44	33.37	53.57	-
	FCL [36]	29.12	23.67	30.75	31.31	25.62	33.02	52.35	-
	SG2HOI [8]	20.93	18.24	21.78	24.83	20.52	25.32	53.30	-
One Stage	UnionDet [12]	17.58	11.72	19.33	19.76	14.68	21.27	47.50	56.20
	IP-Net [37]	19.56	12.79	21.58	22.05	15.77	23.92	51.00	-
	HOI-Trans [14]	26.61	19.15	28.84	29.13	20.98	31.57	52.90	-
	ASNet [38]	28.87	24.25	30.25	31.74	27.07	33.14	53.90	-
	GGNet [13]	23.47	16.48	25.60	27.36	20.23	29.48	54.70	-
	HOTR [15]	25.10	17.34	27.42	-	-	-	55.20	**64.40**
	PhraseHOI(R50) [39]	29.29	22.03	31.46	31.97	23.99	34.36	57.40	-
	QPIC(R50) [4]	29.08	22.48	31.05	31.41	24.00	33.63	58.60	60.90
	QPIC(R101) [4]	29.64	23.27	31.55	32.32	26.21	34.15	58.30	60.70
	Ours(R50)	29.69	22.62	31.81	31.90	24.38	34.14	59.31	61.30
	Ours(R101)	**30.08**	**24.00**	**31.89**	**32.40**	**26.49**	**34.16**	**60.25**	62.29

where, the $z(f,i)$ denotes taking f the i^{th}. $c_i^{\mathcal{P}_j}$ denotes the classification result of i in the j^{th} path, $i \in \{o, inter\}$ and $j \in \{0,1\}$. δ denotes the dual-way matching set and $j,k \in \delta$. Regarding the consistency of box regression, we adopt the Mean-Square Error (MSE) to measure the consistency between matched sets.

$$\mathcal{L}_{b_i} = \frac{1}{2\mathcal{N}} \sum_{j,k\in\delta, i\in(h,o)} \left\| z(box_h^{\mathcal{P}_1}, j) - z(box_h^{\mathcal{P}_2}, k) \right\|_1 \quad (14)$$

where, the $box_i^{\mathcal{P}_j}$ in the formula denotes the box regression result for i in the j^{th} path. The $\mathcal{N}$ denotes the total number of sets in the ground truth. Thus, we can obtain the total consistency loss function as follows:

$$\mathcal{L}_{con} = \lambda_1 \mathcal{L}_{c_o} + \lambda_2 \mathcal{L}_{c_{inter}} + \lambda_3 \sum_{i\in(h,o)} \mathcal{L}_{b_i} \quad (15)$$

where λ_1, λ_2 and λ_3 are the hyper-parameters for adjusting the weights of each loss. We set 1, 0.5 and 1, respectively.

E. Evolution

Training. During the training, both of our paths are matched with the prediction results using bipartition matching [40] with ground truth like the previous work on single stream [4, 14, 15] to get a one-to-one result. Then, the losses are calculated, which consist of three kinds of losses on each path: box regression L1 loss ($\mathcal{L}_i$), GIoU loss [41] ($\mathcal{L}_k$) and class of cross-energy loss and focal loss [42] ($\mathcal{L}_l$). So we can obtain the total losses for both paths as follows:

$$\begin{aligned}\mathcal{L}_{hoi} \quad = & \sum_{i\in(\mathcal{P}_1,\mathcal{P}_2)} (\lambda_b \sum_{j\in(h,o)} \mathcal{L}_{b_j}^i + \lambda_u \sum_{k\in(h,o)} \mathcal{L}_{u_k}^i \\ & + \sum_{l\in(o,inter)} \lambda_c^l \mathcal{L}_{c_l}^i) \end{aligned} \quad (16)$$

where, $\mathcal{L}_{b_i}^j$ in the formula denotes L1 loss of j in i^{th} path. $\mathcal{L}_k$ denotes the IoU loss of k in i^{th} path. $\mathcal{L}_l$ denotes the class cross-energy loss of k in i^{th} path. λ_b, λ_u and λ_c are hyper-parameters, which we set to 2.5, 1 and 1 respectively. Combining the formula, we obtain our total loss:

$$\mathcal{L} = \mathcal{L}_{hoi} + w(t)\mathcal{L}_{con} \quad (17)$$

where, we use the linear ramp-up [26, 43] function $w(t)$. Because at lower training epochs, $\mathcal{P}_1$ and $\mathcal{P}_2$ are inaccurate for HOI detection. Too large a proportion of consistency loss will cause it to focus on negative. So we adopt the ramp-up function to gradually increase the proportion of consistency loss as the epochs of training increase.

Inference. During the inference, we keep $\mathcal{P}_1$ and remove the prediction of $\mathcal{P}_2$. So only the prediction of $\mathcal{P}_1$ is used for bipartite matching, which ensures speed of inference.

IV. Experiments

A. Dataset and Metric

Datasets. We evaluate our model on two large-scale HOI Detection benchmarks: **HICO-DET** [44] and **V-COCO** [45]. HICO-DET has $38,118$ for training and $9,658$ for testing. There are 600 classes of HOI triplets comprising 80 objects and 117 interactions. V-COCO is a subset of COCO [46] and contains $2,533$ for training, $2,867$ for validating, and $4,946$ for testing, which includes 29 interaction categories (4 body motions without interaction with any object) and 80 objects same as COCO. There are 263 classes of HOI triplets.

Evaluation Metric. In HICO-DET [44], we use mAP (mean Average Precision) following the settings referred in [44]. True positive of HOI triplet prediction is required when bounding boxes of human and object between predictions and ground truth have IoU larger than 0.5. Besides, the HICO-DET had been divided into three different parts: Full (600 HOIs), Rare (138 HOIs), Non-Rare (362 HOIs). **In V-COCO** [45], we use AP (Average Precision). There are two evaluation scenarios: scenario 1 contains the 29 actions with 4 body motions and scenario 2 includes the 25 actions, ignoring no-objects HOI categories.

B. Implementation Details

We optimize our network with AdamW. Its weight decay is $1e-4$. The query size is set to 100. In HICO-DET, We initialize our network using the DETR [18] trained in MS-COCO [46]. In V-COCO, the dataset for training DETR excludes the images in V-COCO [45]. All our experiments are trained for 90 epochs and decay learning rate by 10 times at 60^{th} epoch. We conduct our experiments on three NVIDIA GeForce RTX 3090 GPUS with batchsize of 12.

Due to our devices being so different from the official ones, we re-implement the result of QPIC on V-COCO and HICO-DET. Our training strategy and loss weight coefficient are both identical to settings reported in [4].

C. Results and analysis

We compare the performance of our network with the state-of-the-art model on HICO-DET and V-COCO, respectively. Methods for HOI Detection are divided into one-stage and two-stage.

Results on HICO-DET. In Table II, our network makes relative improvements of 0.61, 0.13 and 0.76 in R50, 0.44, 0.73 and 0.34 in R101 with QPIC [4]. We discover that our network makes major improvements on rare categories with R101. Compared to PharseHOI [39], we improve 0.40, 0.59 and 0.35 in R50. These indicate that our QPIC-based study is valid and performs well compared to existing excellent models.

Results on V-COCO. In Table II, compared to QPIC [4], we improve 0.71 and 0.4 under scenarios 1 and 2 in R50, respectively. Compared to PharseHOI [39], we improve 1.71 under scenario 1 in R50. Only scenario 2 is lower than HOTR [15]. These also demonstrate our model work well on V-COCO.

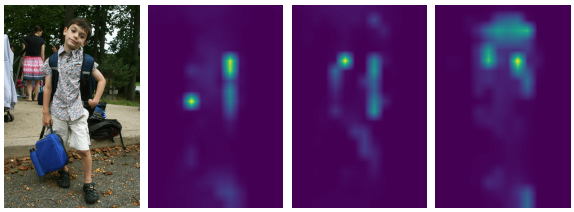
D. Ablation Studies

We do ablation experiments on V-COCO, with the result in Table III. All experiments adopt Resnet50 [28] as backbone.

DPC. In Table III for 1 and 4, we observe mAPs of scenario 1 and 2 rise 0.36 and 0.2. In Table III for 1, 2 and 4, the models, only having $\mathcal{P}_1$ or $\mathcal{P}_2$, are lower than model with DPC (4). In Fig. 7a-7b, we also find attention map more distinct with DPC than QPIC. These mean bringing two paths together and keeping their consistency can help the network improve generalization.

TABLE III: The ablation of network modules. In $\#Param$, the first and second are the parameters for training and inference, respectively.

	Module			Scenario 1	Scenario 2	#Param
	P1	P2	IIA			
1	✓			58.80	61.00	4.14K/4.14K
2		✓		58.35	60.26	4.14K/4.14K
3		✓	✓	58.72	60.81	4.19K/4.19K
4	✓	✓		59.16	61.20	4.14K/4.14K
5	✓	✓	✓	**59.51**	**61.49**	4.19K/4.14K



(a) (b) (c) (d)

Fig. 7: The visualization of ablation. All the attention maps are visualized using the last layer of the decoder. More specific regions and brighter attention maps indicate that noticing features is more effective. (b)-(d) are QPIC, SGPT without IIA, SGPT(Ours).

IIA. In Table III for 4 and 5, ours with IIA improves network mAP 0.71 and 0.49 in V-COCO. In Table III for 2 and 3, we can find that $\mathcal{P}_2$ with IIA improves 0.37 and 0.51. In Fig. 7a-7c, IIA enables the attention map to focus on more details while enhancing interaction areas. These show the importance of IIA in improving $\mathcal{P}_2$ and helping $\mathcal{P}_1$ learn more robust features.

In addition, our network only reaches 4.19K parameters during training, while parameters are consistent with QPIC during inference. This indicates that our inference model improves the accuracy without increasing parameters.

V. Conclusion

In this paper, we proposed a novel method, using the Secondary path ($\mathcal{P}_2$) Guided the Primary Path ($\mathcal{P}_1$) in Transformers, named SGPT. The SGPT contained two core modules: the dual-path consistency (DPC) module and the instance interaction attention (IIA) module. Experiments showed DPC and IIA could make the $\mathcal{P}_2$ enhanced and enable $\mathcal{P}_2$ to better induce the $\mathcal{P}_1$. Abound experiments were conducted on both V-COCO and HICO-DET datasets. The experimental results illustrated that our method achieved advanced performance. The ablation studies proved the effectiveness of the modules (DPC and IIA) proposed in our model. In the future, we will pay attention to improving the generalization and promoting the development of robot manipulation and behavior monitoring.

ACKNOWLEDGMENT

This work is partially supported by the National Natural Science Foundation of China under Grant (No.61906168, No.U20A20196, No.62272267, No.61976191); Zhejiang Provincial Natural Science Foundation of China (Grant No.LY23F020023); Construction of Hubei Provincial Key Laboratory for Intelligent Visual Monitoring of Hydropower Projects (Grant No. 2022SDSJ01) and Hangzhou AI major scientific and technological innovation project (2022AIZD0061).

References

[1] A. Belardinelli, A. R. Kondapally, D. Ruiken, D. Tanneberg, and T. Watabe, "Intention estimation from gaze and motion features for human-robot shared-control object manipulation," 2022.

[2] K. Muvvala, P. Amorese, and M. Lahijanian, "Let's collaborate: Regret-based reactive synthesis for robotic manipulation," in 2022 International Conference on Robotics and Automation, ICRA 2022, Philadelphia, PA, USA, May 23-27, 2022. IEEE, 2022, pp. 4340–4346.

[3] R. Peddi and N. Bezzo, "Interpretable run-time prediction and planning in co-robotic environments," in IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2021, Prague, Czech Republic, September 27 - Oct. 1, 2021. IEEE, 2021, pp. 2504–2510.

[4] M. Tamura, H. Ohashi, and T. Yoshinaga, "QPIC: query-based pairwise human-object interaction detection with image-wide contextual information," in IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021. Computer Vision Foundation / IEEE, 2021, pp. 10 410–10 419.

[5] H. Wang, W. Zheng, and Y. Ling, "Contextual heterogeneous graph network for human-object interaction detection," in Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XVII, ser. Lecture Notes in Computer Science, A. Vedaldi, H. Bischof, T. Brox, and J. Frahm, Eds., vol. 12362. Springer, 2020, pp. 248–264.

[6] X. Zhong, C. Ding, X. Qu, and D. Tao, "Polysemy deciphering network for robust human-object interaction detection," Int. J. Comput. Vis., vol. 129, no. 6, pp. 1910–1929, 2021.

[7] C. Gao, J. Xu, Y. Zou, and J. Huang, "DRG: dual relation graph for human-object interaction detection," in Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XII, ser. Lecture Notes in Computer Science, A. Vedaldi, H. Bischof, T. Brox, and J. Frahm, Eds., vol. 12357. Springer, 2020, pp. 696–712.

[8] T. He, L. Gao, J. Song, and Y. Li, "Exploiting scene graphs for human-object interaction detection," in 2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021. IEEE, 2021, pp. 15 964–15 973.

[9] F. Z. Zhang, D. Campbell, and S. Gould, "Spatially conditioned graphs for detecting human-object interactions," in 2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021. IEEE, 2021, pp. 13 299–13 307.

[10] S. Ren, K. He, R. B. Girshick, and J. Sun, "Faster R-CNN: towards real-time object detection with region proposal networks," in Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada, C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, Eds., 2015, pp. 91–99.

[11] M. Yang, M. Zhou, Z. Li, J. Liu, L. Pan, H. Xiong, and I. King, "Hyperbolic graph neural networks: A review of methods and applications," CoRR, vol. abs/2202.13852, 2022.

[12] B. Kim, T. Choi, J. Kang, and H. J. Kim, "Uniondet: Union-level detector towards real-time human-object interaction detection," in Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XV, ser. Lecture Notes in Computer Science, A. Vedaldi, H. Bischof, T. Brox, and J. Frahm, Eds., vol. 12360. Springer, 2020, pp. 498–514.

[13] X. Zhong, X. Qu, C. Ding, and D. Tao, "Glance and gaze: Inferring action-aware points for one-stage human-object interaction detection," in IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021. Computer Vision Foundation / IEEE, 2021, pp. 13 234–13 243.

[14] C. Zou, B. Wang, Y. Hu, J. Liu, Q. Wu, Y. Zhao, B. Li, C. Zhang, C. Zhang, Y. Wei, and J. Sun, "End-to-end human object interaction detection with HOI transformer," in IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021. Computer Vision Foundation / IEEE, 2021, pp. 11 825–11 834.

[15] B. Kim, J. Lee, J. Kang, E. Kim, and H. J. Kim, "HOTR: end-to-end human-object interaction detection with transformers," in IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021. Computer Vision Foundation / IEEE, 2021, pp. 74–83.

[16] Y. Liao, S. Liu, F. Wang, Y. Chen, C. Qian, and J. Feng, "PPDM: parallel point detection and matching for real-time human-object interaction detection," in 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020. Computer Vision Foundation / IEEE, 2020, pp. 479–487.

[17] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021. OpenReview.net, 2021.

[18] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part I, ser. Lecture Notes in Computer Science, A. Vedaldi, H. Bischof, T. Brox, and J. Frahm, Eds., vol. 12346. Springer, 2020, pp. 213–229.

[19] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in 2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021. IEEE, 2021, pp. 9992–10 002.

[20] A. S. M. Iftekhar, S. Kumar, R. A. McEver, S. You, and B. S. Manjunath, "Gtnet: Guided transformer network for detecting human-object interactions," CoRR, vol. abs/2108.00596, 2021.

[21] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, Eds., 2014, pp. 568–576.

[22] D. Wettschereck and D. W. Aha, "Weighting features," in Case-Based Reasoning Research and Development, First International Conference, ICCBR-95, Sesimbra, Portugal, October 23-26, 1995, Proceedings, ser. Lecture Notes in Computer Science, M. M. Veloso and A. Aamodt, Eds., vol. 1010. Springer, 1995, pp. 347–358.

[23] T. Hastie, R. Tibshirani, and J. H. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd Edition, ser. Springer Series in Statistics. Springer, 2009.

[24] G. Wang, T. Zhang, Y. Yang, J. Cheng, J. Chang, X. Liang, and Z. Hou, "Cross-modality paired-images generation for rgb-infrared person re-identification," in The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020. AAAI Press, 2020, pp. 12 144–12 151.

[25] J. Jeong, S. Lee, J. Kim, and N. Kwak, "Consistency-based semi-supervised learning for object detection," in Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada, H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. B. Fox, and R. Garnett, Eds., 2019, pp. 10 758–10 767.

[26] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," in Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA, I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, and R. Garnett, Eds., 2017, pp. 1195–1204.

[27] T. Miyato, S. Maeda, M. Koyama, and S. Ishii, "Virtual adversarial training: A regularization method for supervised and semi-supervised learning," IEEE Trans. Pattern Anal. Mach. Intell., vol. 41, no. 8, pp. 1979–1993, 2019.

[28] R. Ferjaoui, M. A. Cherni, F. Abidi, and A. Zidi, "Deep residual learning based on resnet50 for COVID-19 recognition in lung CT images," in 8th International Conference on Control, Decision and Information Technologies, CoDIT 2022, Istanbul, Turkey, May 17-20, 2022. IEEE, 2022, pp. 407–412.

[29] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA, I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, and R. Garnett, Eds., 2017, pp. 5998–6008.

[30] P. Gao, M. Zheng, X. Wang, J. Dai, and H. Li, "Fast convergence of DETR with spatially modulated co-attention," in 2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021. IEEE, 2021, pp. 3601–3610.

[31] S. Woo, J. Park, J. Lee, and I. S. Kweon, "CBAM: convolutional block attention module," in Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part VII, ser. Lecture Notes in Computer Science, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds., vol. 11211. Springer, 2018, pp. 3–19.

[32] Y. Liu, Q. Chen, and A. Zisserman, "Amplifying key cues for human-object-interaction detection," in Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XIV, ser. Lecture Notes in Computer Science, A. Vedaldi, H. Bischof, T. Brox, and J. Frahm, Eds., vol. 12359. Springer, 2020, pp. 248–265.

[33] Z. Hou, X. Peng, Y. Qiao, and D. Tao, "Visual compositional learning for human-object interaction detection," in Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XV, ser. Lecture Notes in Computer Science, A. Vedaldi, H. Bischof, T. Brox, and J. Frahm, Eds., vol. 12360. Springer, 2020, pp. 584–600.

[34] D. Kim, X. Sun, J. Choi, S. Lin, and I. S. Kweon, "Detecting human-object interactions with action co-occurrence priors," in Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XXI, ser. Lecture Notes in Computer Science, A. Vedaldi, H. Bischof, T. Brox, and J. Frahm, Eds., vol. 12366. Springer, 2020, pp. 718–736.

[35] D. Jin, X. Ma, C. Zhang, Y. Zhou, J. Tao, M. Zhang, and Z. Li, "Towards overcoming false positives in visual relationship detection," in 32nd British Machine Vision Conference 2021, BMVC 2021, Online, November 22-25, 2021. BMVA Press, 2021, p. 248.

[36] Z. Hou, B. Yu, Y. Qiao, X. Peng, and D. Tao, "Detecting human-object interaction via fabricated compositional learning," in IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021. Computer Vision Foundation / IEEE, 2021, pp. 14 646–14 655.

[37] T. Wang, T. Yang, M. Danelljan, F. S. Khan, X. Zhang, and J. Sun, "Learning human-object interaction detection using interaction points," in 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020. Computer Vision Foundation / IEEE, 2020, pp. 4115–4124.

[38] M. Chen, Y. Liao, S. Liu, Z. Chen, F. Wang, and C. Qian, "Reformulating HOI detection as adaptive set prediction," in IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021. Computer Vision Foundation / IEEE, 2021, pp. 9004–9013.

[39] Z. Li, C. Zou, Y. Zhao, B. Li, and S. Zhong, "Improving human-object interaction detection via phrase learning and label composition," in Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022. AAAI Press, 2022, pp. 1509–1517.

[40] H. W. Kuhn, "The hungarian method for the assignment problem," in 50 Years of Integer Programming 1958-2008 - From the Early Years to the State-of-the-Art, M. Jünger, T. M. Liebling, D. Naddef, G. L. Nemhauser, W. R. Pulleyblank, G. Reinelt, G. Rinaldi, and L. A. Wolsey, Eds. Springer, 2010, pp. 29–47.

[41] H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. D. Reid, and S. Savarese, "Generalized intersection over union: A metric and a loss for bounding box regression," in IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019. Computer Vision Foundation / IEEE, 2019, pp. 658–666.

[42] T. Lin, P. Goyal, R. B. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017. IEEE Computer Society, 2017, pp. 2999–3007.

[43] A. Tarvainen and H. Valpola, "Weight-averaged consistency targets improve semi-supervised deep learning results," CoRR, vol. abs/1703.01780, 2017.

[44] Y.-W. Chao, Y. Liu, X. Liu, H. Zeng, and J. Deng, "Learning to detect human-object interactions," in Proceedings of the IEEE Winter Conference on Applications of Computer Vision, 2018.

[45] S. Gupta and J. Malik, "Visual semantic role labeling," arXiv preprint arXiv:1505.04474, 2015.

[46] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in Computer Vision–ECCV 2014. Springer, 2014, pp. 740–755.