Exploring Pose-Aware Human-Object Interaction via Hybrid Learning

Eastman Z Y, Wu[1,2] Yali Li[1,2] Yuan Wang[1,2] Shengjin Wang[1,2*]
[1] Department of Electronic Engineering, Tsinghua University
[2] Beijing National Research Center for Information Science and Technology (BNRist), China
{wu-zy23, wy23}@mails.tsinghua.edu.cn, {liyali13, wgsgj}@tsinghua.edu.cn,

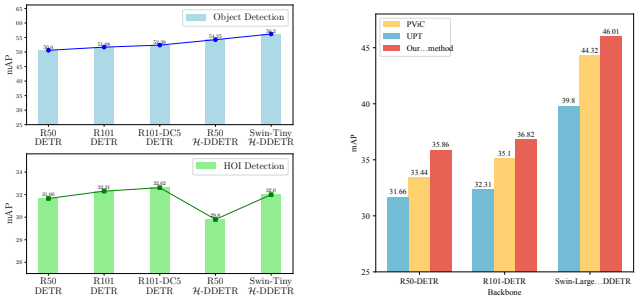
Abstract

Human-Object Interaction (HOI) detection plays a crucial role in visual scene comprehension. In recent advancements, two-stage detectors have taken a prominent position. However, they are encumbered by two primary challenges. First, the misalignment between feature representation and relation reasoning gives rise to a deficiency in discriminative features crucial for interaction detection. Second, due to sparse annotation, the second-stage interaction head generates numerous candidate <human, object> pairs, with only a small fraction receiving supervision. Towards these issues, we propose a hybrid learning method based on pose-aware HOI feature refinement. Specifically, we devise pose-aware feature refinement that encodes spatial features by considering human body pose characteristics. It can direct attention towards key regions, ultimately offering a wealth of fine-grained features imperative for HOI detection. Further, we introduce a hybrid learning method that combines HOI triplets with probabilistic soft labels supervision, which is regenerated from decoupled verb-object pairs. This method explores the implicit connections between the interactions, enhancing model generalization without requiring additional data. Our method establishes state-of-the-art performance on HICO-DET benchmark and excels notably in detecting rare HOIs.

1. Introduction

Human-Object Interaction (HOI) detection, as a significant computer vision task, endeavors to locate the human-object pair <*human, object*> and identify the interactive relationships between them, which unveils the mechanism of how people interact with objects. HOI detection harbors substantial potential in *defacto* applications, *e.g.*, human-computer interaction, AR/VR, video surveillance.

HOI detectors can be categorized into one-stage and two-stage methods based on their architecture. In the context of one-stage methods, achieving model convergence presents a formidable challenge. Due to the sparse supervision provided by triplet labels and the joint training of object detector and interaction head, model convergence typically necessitates hundreds of GPU hours. In contrast, two-stage methods capitalize on object detectors with pre-trained weights, which focuses exclusively on training the interaction head and leads to expedited convergence. Further, two-stage detectors traverse all potential human-object interaction pairs, showcasing unparalleled flexibility in inferring any specified interaction pair within a visual scene.



(a) Object *vs* HOI detetion (b) HOI detection performance

Figure 1. (a) illustrates the Object Detection performance and HOI Detection performance based on DETR[2] and more advanced $\mathcal{H}$-DDETR [15, 49]. (b) depicts the performance enhancement of our method compared to the previous state-of-the-art[45, 46].

Due to the aforementioned advantages, researchers have increasingly shifted their focus towards two-stage architectures. However, certain challenges impede the further progress of two-stage detectors. Through investigative study, we noted that two-stage methods exhibit performance degradation induced by the inter-task gap between object detection and HOI detection. As shown in Figure 1(a), we showcase the performance of detectors in the object detection task and their HOI detection performance after adding the same second-stage interaction head in UPT[45]. The advanced $\mathcal{H}$-DETR outperforms DETR by a remarkable 3.65 mAP in object detection, yet it exhibits the lowest perfor-

*corresponding author

mance in HOI detection. This highlights the degradation of the model's capability with two-stage methods. The object detector inherently incurs task bias inclination, acquiring features well-suited for object detection but overlooking the nuanced feature intricacies necessary for HOI. The previous second-stage interaction heads lack the ability to refine features and extract crucial details. On the other hand, in two-stage methods, the interaction head combines all detected human and object, resulting in a substantial number of candidate interaction pairs. However, only a small portion of pairs are labeled for supervision. In previous paradigms, indiscriminately categorizing the remaining interaction pairs as negative samples would introduces undesirable noise into the training process. This situation constrains the model's generalization capability, making it challenging to achieve effective detection for rare HOIs.

To address the task bias in features, we introduce a pose-aware HOI feature refinement strategy. In previous studies [32, 43, 45, 50], human and object features were treated equally as uniform queries. However, in real life, humans observe with their eyes, hold and grasp objects with their hands, and perform actions like jumping and standing with their feet. This indicates that features related to human pose and relative spatial information are often crucial for detecting human-object interactions. Motivated by this insight, we meticulously designed our strategy to leverage human pose information. Specifically, we estimate human body keypoints and then adaptively generate feature regions for different body parts. We perform detailed pairwise spatial encoding, utilizing criteria such as the intersection over Body-part area (IoB) between human body parts and bounding boxes of detected objects, as well as geometrical information from keypoints. This information is finely utilized to guide attention, and subsequently regenerate fine-grained HOI features from refined object queries. By introducing the Pose-Aware Feature Refinement strategy, our model demonstrates its capability to direct its attention towards information-rich regions, thereby boosting the performance of HOI detection. This will be examined and visually presented in the upcoming ablation studies.

In response to the limitation of detectors supervised by sparse samples, we introduce Hybrid Learning. Leveraging the flexibility of the two-stage approach, we utilize a fully trained interaction head to generate probabilistic soft labels for potential interaction pairs. In contrast to previous studies [40, 42] that constructed artificially annotated additional datasets using diffusion models[31, 33] or language models[17, 30], we opted not to use additional data, but rather to deeply explore the correlations among interactions. For example, "*ride*" and "*straddle*" exhibit positive correlation, while "*stand on*" and "*stand under*" conflict in spatial features. Using probabilistic soft labels allows for the representation of implicit interactions in a probability distribution form, enabling the model to learn decisive representations and enhancing its generalization capabilities.

Our main contribution can be summarized as follows:

- We propose a Pose-Aware Feature Refinement to incorporate intricate spatial features between pose and object, effectively mitigating the task bias within the queries.
- We devise a novel hybrid learning method for HOI detection task, addressing the sparsity of triplet annotations, thereby enhancing model's generalization without requiring additional data.
- Through extensive experimentation, our model achieves the state-of-the-art performance, outperforming all existing HOI detectors on widely-used datasets.

In particular, on the HICO-DET dataset, we attained 35.86 mAP and 36.73 mAP using ResNet50 and ResNet101 as backbone, respectively. Our approach notably reached 46.01 mAP using the Swin-Large as backbone, effectively overcoming the challenge of detecting rare HOI classes and delivering a remarkable performance gain of +2.13 mAP.

2. Related Work

One-stage Methods. One-stage detectors aim to detect HOI triplets in a single forward pass. These methods often leverage predefined interaction points or anchors. [7, 16, 20, 37, 48]. For instance, PPDM[20] introduced a simple yet effective strategy by considering the midpoint between the human and the object as the interaction point. Alternatively, QAHOI[4] proposed a query-based anchors method, generating reference points through deformable transformer decoder. FGAHOI[27] advanced this concept by generate fine-grained anchors that guide HOI feature extraction from complex tasks. Beyond anchor strategies, some studies have focused on innovating the detection process in one-stage methods. ERNet[22] integrated a predictive uncertainty estimation framework in their classification heads, thus improving prediction robustness, while CDN[43] introduced the advantages of two-stage detectors into one-stage methods by disentangling human-object detection and interaction classification in a cascading manner.

Two-stage Methods. In recent research, the advantages of two-stage methods have been gradually revealed. Dividing the HOI detection task into object detection and interaction reasoning has demonstrated improved efficiency and flexibility. Two-stage methods employ the off-the-shelf object detector[2, 5, 11, 26] to obtain detections, enriching HOI interaction features by effective utilization of object queries or incorporation of additional information, such as spatial [1, 19, 38], pose [6, 10, 13, 18, 34], and graph features [28, 29, 35, 38, 44], along with instance-centric attention mechanisms[8, 36], and linguistic cues[9, 25, 41]. Specifically, iCAN[8] developed the instance-centric attention network, leveraging contextual image features to en-

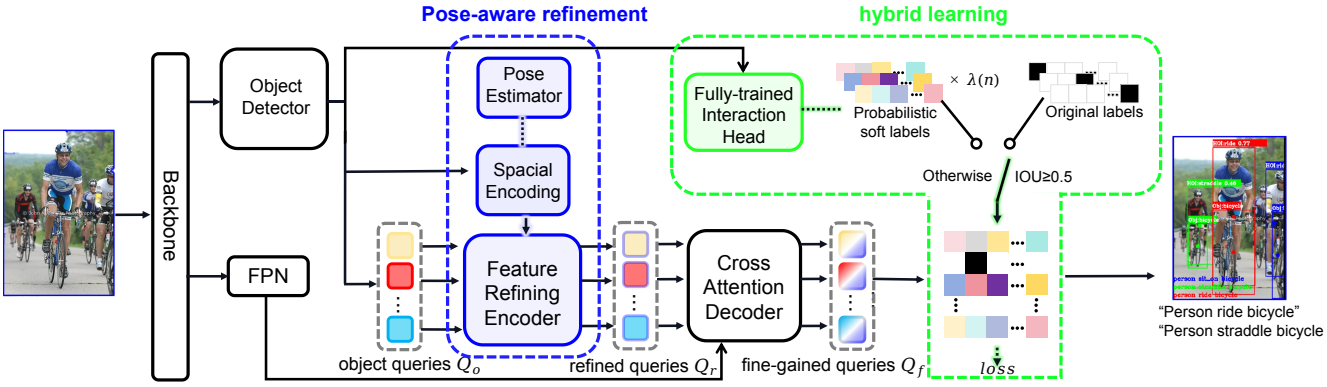


Figure 2. **Overall architecture of our method.** The blue component represents the Pose-Aware Feature Refinement, which fuses human body pose information to refine object queries. The green part represents the Hybrid Learning, generating soft label supervision for H-O pairs. The parameters of the Pose Estimator and the Fully-trained Interaction head are not involved in gradient computation.

hance human-object pair representations. UPT[45] elucidated that using self-attention on unary features and interaction pairs effectively enhances the confidence of positive samples. PViC[46] reintroduced image features into the H-O pairs representation via cross-attention to counteract the lack of relevant contextual information, while RLIPv1[41] incorporated linguistic features to improve few-shot HOI detection. They performed contrastive language-image pre-training and showcased its effectiveness. Additionally, some studies aim to address the issue of sparse annotations by augmenting the dataset[40, 42]. For instance, DiffHOI[40] contributed to data diversity by creating a balanced synthetic dataset, encompassing over 140K images with comprehensive HOI triplet annotations.

3. Method

In this section, we provide a detailed introduction to our method. First, we present the overall architecture, as depicted in Figure 2. Subsequently, we delve into the specifics of our Pose-Aware feature refinement in Section 3.1 and the Hybrid Learning in Section 3.2.

3.1. Pose-Aware Feature Refinement

To distill essential information from queries affected by task bias, we center our focus on human poses. Integrating human body pose features into HOI detection offers evident benefits. First, it refines human queries by fusing the appearance features of the human body, emphasizing the dominant role of humans in interactions. Second, it enables a finer representation of paired spatial features by considering geometric relationships between human keypoints and objects. Building upon this idea, we design a pose-aware spatial encoding to guide the attention focusing on significant regions with rich interactive information, thereby extracting deep representational features suitable for HOI detection.

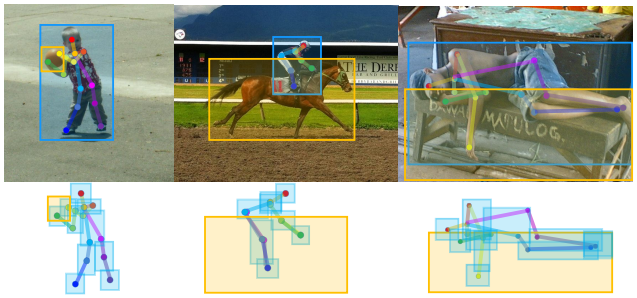


Figure 3. Visualizations. First row illustrates the body keypoints detected by ViTPose[39] and object bounding boxes. Second row shows generated human body-part regions.

Body-part Region Generation. We employ an off-the-shelf pose estimator for detecting human body keypoints, enabling us to dynamically locate body-part regions. More precisely, we define each body-part region by extending the keypoint coordinates along the X and Y axes, using coefficients that adjust based on their proportional sizes. A body-part region is specified by a central keypoint coordinate (x_i, y_i) and an adjacent auxiliary keypoint coordinate (x_j, y_j). The top-left and bottom-right vertices of the generated body-part region's bounding box are determined by $(x_i - \Delta x, y_i - \Delta y)$ and $(x_i + \Delta x, y_i + \Delta y)$, respectively. This means that the regions are centered around (x_i, y_i), with a width of $2 \times \Delta x$ and a height of $2 \times \Delta y$. Here, Δx and Δy represent the adaptive offsets along the X and Y axes, respectively, calculated as indicated in Eq.1.

$$\begin{cases} \Delta x = \left(\alpha + \beta \frac{r_{i,j}}{1+r_{i,j}}\right) \times d_{i,j} \\ \Delta y = \left(\alpha + \beta \frac{1}{1+r_{i,j}}\right) \times d_{i,j} \end{cases} \tag{1}$$

where α, β are fixed coefficients. $r_{i,j}$ and $d_{i,j}$ can be for-

mulated as:

$$r_{i,j} = \frac{|x_i - x_j|}{|y_i - y_j|}, \; d_{i,j} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2} \quad (2)$$

In Eq. 2, $r_{i,j}$ specifies the aspect ratio of the rectangle delineated by the central keypoint i and its auxiliary keypoint j, with $d_{i,j}$ quantifying their Euclidean distance. It is worth noting that, when calculating a specific region, the central and auxiliary keypoints are fixed. For example, the computation of the left hand region employs *LWrist* as the central point and *LElbow* as the auxiliary point. This approach ensures stable and precise localization of body parts across various human pose scenarios, as illustrated in Figure 3.

Spatial Encoding with Human Pose. Following the generation of body-part regions and keypoints, we utilize them to encode the spatial features between Human-Object pairs. To incorporate pose information, we include finer handcrafted features encoded through human keypoints. Specifically, we gauge the contribution of each body part in interactions by calculating the Intersection over Body-part Area (IOB) between all body parts and detected objects. When humans perform actions such as 'hold', 'pull', and 'wield', the IOB of hands significantly increases, whereas during distant interactions such as 'watch' and 'fly', all IOBs tend to approach zero. This indicates that IOBs can capture different behavioral patterns. Additionally, we consider factors such as the center of mass, body angles, relative sizes, and interaction directions to enrich our encoding. Then, we process the encoded spatial features through three fully-connected layers with ReLU activation functions to ensure they align with the same dimension as the object queries.

HOI Feature Refinement Encoder. Humans as the main character in interaction should not be considered as simple queries as other objects. Therefore, we enhance human queries by integrating pose features, followed by employing pose-aware spatial encoding to guide the refinement of HOI queries. In detail, let $X_{pose} \in \mathbb{R}^{n_h \times m}$ denote human queries enhanced by fusing pose feature and $X_{object} \in \mathbb{R}^{n_o \times m}$ represent object queries. n is the number of all detections (humans and objects). m is the dimension of queries. Y represents the spatial encoding. We first concatenate X_{pose}, X_{object}, and duplicate them for n times as $\tilde{X} \in \mathbb{R}^{n \times n \times m}$. Then, we utilize a modified attention mechanism outlined in UPT [45]. This involves computing the value V through element-wise multiplication of $\tilde{X}$ and Y, followed by a fully-connected layer. Then using indices of H-O pairs, we get paired features, denoted as $X \in \mathbb{R}^{n \times n \times 2m}$. The attention map W is computed as:

$$W = \text{softmax}(\text{Linear}[\text{concat}(X, Y)]) \quad (3)$$

Subsequently, the final output is obtained by conducting element-wise multiplication of W and V. Thus, by fusing pose features and utilizing pose-aware spatial encoding, we guide the attention map to refine the queries, obtaining refined Human-Object Interaction features.

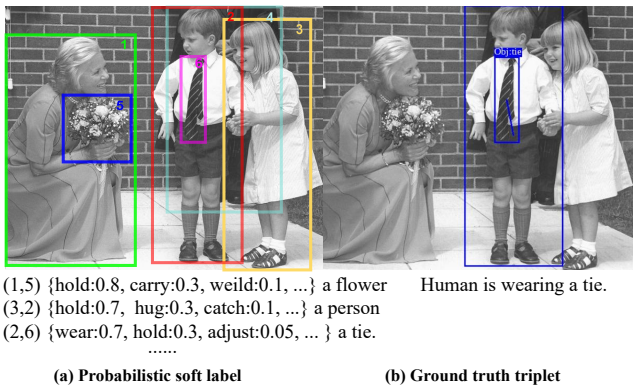


Figure 4. Illustraing our probabilistic soft label and ground truth.

3.2. Hybrid Learning.

The Human-Object Interaction data exhibit a severe long-tail distribution. For example, in the widely-used HICO-DET dataset, there are 138 HOI classes that have fewer than 10 training samples. Moreover, the sparsity in HOI triplet annotation exacerbates the issue, particularly when discerning interactions with subtle variances or reasoning about rare HOIs. One feasible approach is to search for additional data with HOI annotations. However, this approach incurs considerable costs. Instead of labeling new data, we delve deeper into existing data to uncover more information. Considering the inefficiency and ambiguity in generating precise HOI triplets, we predict the probabilities of verbs for each unsupervised human-object interaction pair. This process fundamentally entails the learning of implicit relationships among verbs, represented as a probability distribution. Additionally, it introduces supplementary supervision for rare HOIs through the soft labels generated from decoupled verb-object pairs, thereby reflecting human learning mechanisms: wherein a person who has learned to recognize an action, such as kissing among humans, can readily identify the same action across different species.

Specifically, we initially train a two-stage detector, and then extract the fully-trained interaction head from the detector. During the training phase, the unlabeled Human-Object pairs are fed into the interaction head to generate predictions for the verbs within the interactions. We employ the sigmoid function to convert these scores into probabilities, thus generating soft labels for each H-O pair, which are referred to as $label_s$. We convert the original triplet annotation into a one-hot format, represented as $label_o$. It is important to note that in order to minimize the introduction of inherent biases associated with the interaction head, we modify the $label_s$. Specifically, when there are annotations for H-O pairs in $label_o$, we mask the corresponding values

within $label_s$. Otherwise, we use $label_s$ as supervision. The modified soft labels are denoted as $label_m$, and can be formulated as follows:

$$\begin{cases} label_m^i = \vec{0}, & where\ \ label_o^i \neq \vec{0} \\ label_m^i = label_s^i, & otherwise \end{cases} \tag{4}$$

Eq.5 demonstrates the computation of $\mathcal{L}^{\text{o}}$, the loss associated with the original triplet annotations. In this context, Q represents the set of all possible interaction queries, and n_1 denotes the cardinality of Q, i.e., the number of elements within Q. The procedure initiates by computing the Intersection Over Union (IOU) between the detection boxes for both the human and object in each interaction pair and their corresponding ground truth boxes. If both IOU values exceed 0.5, we proceed to calculate the focal loss using $label_o$.

$$\mathcal{L}^{\text{o}} = \sum_{i}^{n_1} \mathcal{L}_{iou \geq 0.5}^{focal}(Q_i, label_o^i) \tag{5}$$

For computing $\mathcal{L}^{\text{m}}$, the loss for modified probability soft labels, we employ the focal loss between queries and soft labels within the set $\hat{Q}$. Here, $\hat{Q}$ includes all queries lacking Human-Object Interaction triplet supervision, and n_2 indicates the size of $\hat{Q}$. The use of focal loss helps in mitigating the impact of class imbalance by focusing more on challenging, hard-to-classify examples.

$$\mathcal{L}^{\text{m}} = \sum_{i}^{n_2} \mathcal{L}^{\text{focal}}(\hat{Q}_i, label_m^i) \tag{6}$$

In addition, hybrid loss can be expressed as follow, $\mathcal{L}^{\text{h}} = \mathcal{L}^{\text{o}} + \lambda\mathcal{L}^{\text{m}}$. We use the hyper-parameter λ to regulate the extent of hybrid supervision, where a higher λ signifies a firmer endorsement of the generated soft labels. In the default scheme, λ is set to 0.5.

The hybrid learning method enhances the model's generalization without requiring additional data by effectively capturing the underlying correlations between interactions. For instance, for a baseball bat, actions like '*hold*,' '*swing*,' and '*wield*' exhibit positive correlations while actions like '*stand on*' and '*stand under*' showcase negative correlations. From this perspective, the expression of implicit associations between interactions through original HOI triplets is quite limited. In the training process, this might even suppress positively correlated verbs, leading to confusion. In our hybrid learning method, the '*a person is riding a bike*' triplet could be represented as '*confirmed riding*', '*possible sitting on*' '*slightly against standing*' '*against flying*' and so forth. This is equivalent to a collection of triplet labels with weighted representations.

More Variants of Hybrid Learning. We designed several different variants of our hybrid learning process.

Hybrid decay scheme. Different from the default scheme, λ is not a constant coefficient, but a parameter decreases with each epoch. It aligns better with prior knowledge. Initially, assigning higher confidence to soft labels allows better learning of implicit relationships between different interactions by the fully-trained interaction head, thereby providing a smoother gradient descent. In the later stages of training, as the model's performance approaches or surpasses that of the interaction head, reducing the confidence in soft labels helps avoid the model being overly influenced by potential errors in the interaction head's understanding. This allows the model to generate alternative interpretations and perceptions.

$$\lambda(t) = \max(\sigma\frac{t_{stop} - t}{t_{stop}}, 0) \tag{7}$$

As represented in Eq. (7), $\lambda(t)$ starts with the value σ and linearly decreases until it reaches 0 at t_{stop}. In conclusion, we compute hybrid loss as $\mathcal{L}^{\text{h}} = \mathcal{L}^{\text{o}} + \lambda(t)\mathcal{L}^{\text{m}}$ with decaying $\lambda(t)$ in the first t_{stop} epochs and $\mathcal{L}^{o}$ only after t_{stop} epochs.

Hybrid layer scheme. Inspired by [15], we devised a hybrid layer scheme. In this scheme, distinct supervision strategies are employed for different decoder layers. Hybrid supervision targets lower-level decoder layers, whereas high-level layers use only $label_o$ for loss calculation. Hybrid loss is shown in Eq.(8).

$$\mathcal{L}^{\text{h}} = \sum_{i=n_l+1}^{n} \mathcal{L}_i^{\text{o}} + \sum_{j=0}^{n_l} \mathcal{L}_j^{\text{o}} + \lambda\mathcal{L}_j^{\text{m}} = \sum_{i=0}^{n} \mathcal{L}_i^{\text{o}} + \sum_{j=0}^{n_l} \lambda\mathcal{L}_j^{\text{m}} \tag{8}$$

4. Experiments

This section covers the experiments from four perspectives: experimental settings, performance comparison with state-of-the-art methods, module effectiveness via ablation studies, and qualitative results with a discussion on limitations.

4.1. Experimental Settings

Dataset and Evaluation We train and test our model on two common datasets: HICO-DET [3] and V-COCO [12].

HICO-DET comprises 47,776 images. It includes 80 object categories and 117 verb classes, resulting in 600 types of HOI triplets. HICO-DET consists of three subsets: (i) Full, comprising all 600 HOI triplets. (ii) Rare, encompassing 138 HOI triplets with fewer than 10 training samples. (iii) Non-Rare, other 462 HOI triplets. V-COCO is a subset of MS-COCO [23], with a smaller scale compared to HICO-DET. It comprises 10,346 images. V-COCO contains 24 interactions and 80 objects. We follow previous approach [3, 12, 45] to evaluate our model on HICO-DET and VCOCO. Specifically, for HICO-DET, we conducted evaluations under both the Default setting and the Known

Method	Backbone	HICO-DET						V-COCO	
		Default			Known Object			Default	
		Full	Rare	Non-rare	Full	Rare	Non-rare	AP^{S1}_{role}	AP^{S2}_{role}
FCL [14]	ResNet-50	24.68	20.03	26.07	26.80	21.61	28.35	52.4	-
QPIC [32]		29.07	21.85	31.23	31.68	24.14	33.93	58.8	61.0
UPT [45]		31.66	25.94	33.36	35.05	29.27	36.77	59.0	64.5
STIP [47]		32.22	28.15	33.43	35.29	31.43	36.45	**66.0**	**70.7**
$DiffHOI_s$ [40]		34.41	31.07	35.40	37.31	34.56	38.14	61.1	63.5
PViC [46]		34.69	32.14	35.45	38.14	35.38	38.97	59.7	65.4
$Ours_s$		**35.86**	**32.48**	**36.86**	**39.48**	**36.10**	**40.49**	61.1	66.6
HOITrans [51]	ResNet-101	26.61	19.15	28.84	29.13	20.98	31.57	52.9	-
QPIC [32]		29.90	23.92	31.69	32.38	26.06	34.27	58.3	60.7
CDN [43]		32.07	27.19	33.53	34.79	29.48	36.38	**63.9**	65.9
UPT [45]		32.62	28.62	33.81	36.08	31.41	37.47	61.3	67.1
GEN-VLKT [21]		34.95	31.18	36.08	38.22	34.36	39.37	63.6	65.9
$Ours_m$		**36.82**	**33.99**	**37.66**	**40.56**	**37.02**	**41.69**	62.3	**68.2**
QAHOI [4]	Swin-Large	35.78	29.80	37.56	37.59	31.36	39.36	-	-
FGAHOI [27]		37.18	30.71	39.11	38.93	31.93	41.02	-	-
$DiffHOI_l$ [40]		40.63	38.10	41.38	43.14	40.24	44.01	**65.7**	68.2
PViC [46]		44.32	44.61	44.24	47.81	48.38	47.64	61.7	68.0
$Ours_l$		**46.01**	**46.74**	**45.80**	**49.50**	**50.59**	**49.18**	63.0	**68.7**

Table 1. Comparison of the our method with current remarkable studies on the HICO-DET and V-COCO datasets. **Bold** and underline items represent the best and the second best one. The PViC's code for the V-COCO is being cleaned up, so we use our reproduced results.

Object setting for the Full, Rare, and Non-Rare subsets. For V-COCO, we evaluated performance in AP^{S1}_{role} and AP^{S2}_{role}.

Implementation Details. For our hybrid learning, in the default scheme, λ is fixed at 0.5 as a constant. In the hybrid decay scheme, σ is set to 1, and n_{stop} is configured as 25, following Equation (7). For our Pose-Aware Feature Refinement, we use $ViTPose_{base}$ as pose estimator. α and β, as specified in Equation (1) for the Body-part Region Generation algorithm, are set to 0.25 and 0.33, respectively. All models are trained for 30 epochs using the AdamW optimizer and a multi-step learning rate decay mechanism. We use focal loss[24] for both $label_o$ and $label_m$. During inference, we multiply the object confidences of H-O pairs and the interaction scores, following previous practices[45, 46].

4.2. Comparison to State-of-the-Art

Table 1 compares our method with previous state-of-the-art methods on the HICO-DET and V-COCO datasets. Approaches are categorized into ResNet backbone and Swin Transformer backbone methods. We report our model's performance with three different scale backbones to demonstrate its scalability and facilitate direct comparison with previous research. Our method outperforms all existing one-stage and two-stage approaches, achieving state-of-the-art performance across both default and Known Object settings on the HICO-DET. Achieving a 46.01 mAP, surpassing recent SOTA, PViC [46], by 1.69 mAP and $DiffHOI_l$ [40] by 5.38 mAP, which was trained using a large number of additional high-quality annotated images. This highlights our method's superior refining capabilities and efficient learning. Our smallest model, based on ResNet50, attains a 35.86 mAP, marking a 3.4% and 13.3% improvement over the latest and prior SOTA methods, respectively.

Notably, our model achieves an impressive 46.74 mAP on the rare HOIs subset, which has fewer than 10 training samples for every interaction class. This represents a 2.13 mAP increase and a 4.8% relative improvement compared to the most recent SOTA. In the listed methods, the average detection performance for rare classes is 4.5 mAP lower than for the full classes. However, our approach has significantly narrowed this gap, achieving a performance that is even 0.73 mAP higher than for the full classes with our largest model. This success can be attributed to the effective label enrichment by the hybrid learning paradigm.

4.3. Ablation Study

In this section, we perform ablation studies to assess and analyze the effectiveness of the proposed modules

Comparing different hybrid training schemes. We evaluated the performance of HOI detection across four distinct

#	Hybrid Scheme			Default			Known Object		
	default	decay	layer	Full	Rare	N-rare	Full	Rare	N-rare
A_1				33.44	28.89	34.80	37.38	33.33	38.59
A_2	✓			$34.06^{+0.6}$	$29.96^{+1.1}$	$35.29^{+0.5}$	$37.95^{\mathbf{+0.6}}$	$33.67^{+0.3}$	$39.23^{\mathbf{+0.6}}$
A_3		✓		$34.42^{\mathbf{+1.0}}$	$31.13^{+2.2}$	$35.41^{\mathbf{+0.6}}$	$37.85^{+0.5}$	$34.61^{+1.3}$	$38.81^{+0.2}$
A_4			✓	$34.03^{+0.6}$	$29.64^{+0.8}$	$35.34^{+0.5}$	$37.90^{+0.5}$	$33.65^{+0.3}$	$39.11^{+0.5}$
A_5		✓	✓	$34.25^{+0.8}$	$31.37^{\mathbf{+2.5}}$	$35.11^{+0.3}$	$37.88^{+0.5}$	$35.03^{\mathbf{+1.7}}$	$38.73^{+0.1}$

Table 2. The mAP (%) performance of our proposed method with different hybrid training scheme on the HICO-DET test set.

hybrid schemes, employing DETR with ResNet50 for detection. The detailed results are summarized in Table 2. It is clear that hybrid learning consistently surpasses the baseline in all four schemes, achieving this without any additional data and requiring only a modest increase in training time. The schemes exhibit an average improvement of 0.75 mAP over the baseline. When examining the performance across the full, rare, and non-rare categories, the average improvements are 0.75 mAP, 1.65 mAP, and 0.48 mAP, respectively. Notably, hybrid learning significantly boosts the detection performance of rare HOIs.

When comparing these four schemes specifically, the model performs best under the decay mode, showing an improvement of 1.0 mAP. The other three schemes exhibit similar performance improvements, each achieving an increase of approximately 0.7 mAP. As mentioned earlier, this could be attributed to better alignment with the prior assumptions. In the default scheme, using constant weights introduces noise when the model's performance surpasses that of the fully-trained interaction head. In contrast, in the decay scheme, the confidence weight $\lambda(t)$ of the soft labels gradually decreases with iterations. This approach guides a smoother gradient descent in the early stages and avoids the introduction of noise from misperceptions by the interaction head in the later stages of training.

Hybrid learning on different models. Additionally, we employed various models to validate the effectiveness of the hybrid training module. We selected two-stage HOI detectors: PViC[46], UPT[45], and a UPT variant for evaluation.

The experimental results, as shown in Table 3, demonstrate that after applying the hybrid training method, all three models converged faster and achieved improved performance. For instance, H-PVIC with ResNet50 as the backbone reached 33.5 mAP in 19 epochs, while PViC, trained with only HOI triplet supervision, achieved the same performance after 30 epochs. Besides enhancing training efficiency, the hybrid training module has also improved the models' generalization capabilities. After fully training for 30 epochs on the HICO-DET dataset, the performance of H-PViC exceeded that of PViC by 1.0 mAP and 0.9 mAP, using ResNet50 and ResNet101 as backbones, respectively.

Method	Backbone	Epoch	Full	Rare	N-rare
UPT[45]	R50	20	31.6	25.6	33.4
$\mathcal{H}$-UPT	R50	**15**	31.6	26.1	33.3
$\mathcal{H}$-UPT	R50	20	$\mathbf{31.9}^{+0.3}$	26.2	33.6
UPT*	R50	30	30.0	24.1	31.8
$\mathcal{H}$-UPT*	R50	**21**	30.1	24.9	31.6
$\mathcal{H}$-UPT*	R50	30	$\mathbf{30.4}^{+0.4}$	25.3	31.4
PVIC[46]	R50	30	33.4	28.9	34.8
$\mathcal{H}$-PVIC	R50	**19**	33.5	29.6	34.7
$\mathcal{H}$-PVIC	R50	30	$\mathbf{34.4}^{+1.0}$	31.1	35.4
PVIC	R101	30	34.8	31.2	35.6
$\mathcal{H}$-PVIC	R101	**19**	34.9	30.2	36.3
$\mathcal{H}$-PVIC	R101	30	$\mathbf{35.7}^{+0.9}$	30.8	37.0

Table 3. The mAP (%) of different models employ hybrid learning on HICO-DET test set. UPT* refers to a variant of UPT that removed the modified attention transformer decoder. $\mathcal{H}$-model denotes model with hybrid learning.

#	Module		Default Setting (mAP)		
	Hybrid	Pose	Full	Rare	N-rare
A_1			33.44	28.89	34.80
A_3	✓		34.42	31.13	35.41
B_1		✓	35.39	31.51	36.55
B_2	✓	✓	35.86	32.48	36.86

Table 4. Effect of Hybrid learning and Pose-aware Refinement.

It is worth noting that the fully trained interaction head we utilized was only capable of achieving a 33.4 mAP performance. This indicates that through hybrid soft-label supervised training, the model found a more optimal global minimum during gradient descent, improving the model's generalization capabilities without altering the model design.

Pose-Aware Feature Refinement Module. Table 4 explore the effectiveness of our Pose-aware feature Refinement module. Comparing models A_1 and B_1, it is evident that directly incorporating the pose-aware module results in a notable improvement of 1.95 mAP compared to the baseline. We also visualize the attention map in the last decoder

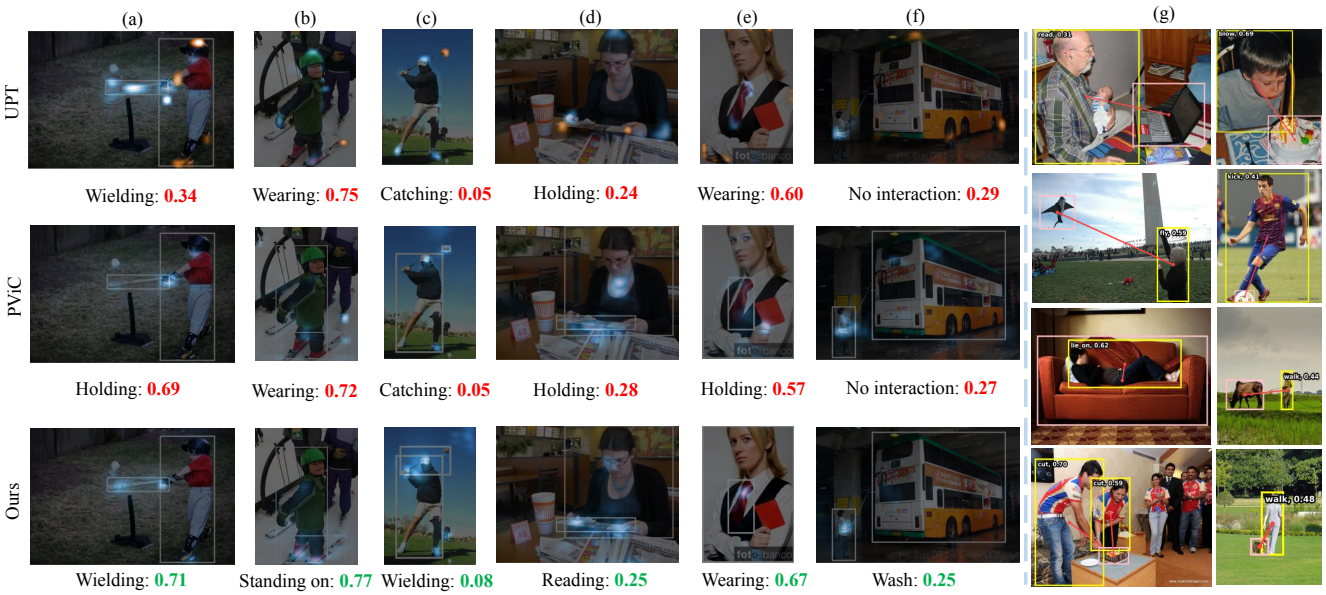


Figure 5. Illustrating the average attention maps for different HOI detectors, the three rows correspond to the visualizations of UPT, PViC, and our method, respectively. For UPT, since it doesn't employ global cross-attention, we visualize the attention map for the queries of human and object in interaction pairs, represented in two colors. Some qualitative results are shown in Figure (g).

layer to facilitate an intuitive interpretation of the impact brought by the pose-aware module, as shown in Figure.5.

As previous study [46] pointed out, the frozen object features extracted from object detector often pool information from the box boundary since this aids localisation. As indicated in the first line, this approach lacks the adaptability to focus on information-rich regions autonomously. Our approach considers the significance of human body limbs engaged in interactions, guiding attention to focus on crucial regions. For instance, in Figure 5(b), our method accurately directs attention to the area near the feet, enabling the correct detection of "*a person standing on a skateboard*". Similarly, in (d), owing to the refined human queries, our model concentrates on both the person's eyes and the hand holding a book, accurately detecting the "*reading*" interaction.

Unlike previous research, which relied on pre-defined interaction points or anchors and introduced noise in long-distance scenes, our approach leverages pose-aware spatial information for each human-object pair. In that case, our HOI detector also performs well in scenarios involving distant interactions, such as flying kites. An example is shown in Figure 5(g), where the model learns distant interaction patterns from spatial encoding.

4.4. Qualitative Results and Limitations

Several qualitative results are illustrated in Figure 5(g). Our model accurately identifies information-rich regions within human-object pairs, producing finely-detailed features. For instance, by concentrating on the facial expressions of a child blowing out a candle and the grip of a person holding a rope, our model precisely detects these interactions. However, as our model is trained with soft label supervision, it tends to encompass a wider array of potential interactions. This approach leads to a decrease in confidence when predicting specific human-object interactions. Moreover, the detector sometimes exhibits overconfidence in frequently occurring actions, such as "*wear*" and "*hold*".

5. Conclusion

In this paper, we analyze the limitations of existing two-stage HOI detection methods, which are constrained by feature task bias and the lack of HOI annotations. To address these issues, we propose the Pose-Aware Feature Refinement, leveraging human body pose information to encode spacial features, guiding attention to refine queries, thus obtain fine-grained HOI interaction features. We further introduce hybrid learning that generates probabilistic soft labels for unsupervised potential H-O pairs, enhancing the model's generalization ability without additional data. We finally achieve exceptional performance, a new state-of-the-art on widely used datasets.

Acknowledgement

This work is supported by the state key development program in 14th Five-Year under Grant No. 2021YFF0602103, 2021YFF0602102, 2021QY1702. We also thank for the research fund under Grant No. 2019GQG0001 from the Institute for Guo Qiang, Tsinghua University.

References

[1] Ankan Bansal, Sai Saketh Rambhatla, Abhinav Shrivastava, and Rama Chellappa. Detecting human-object interactions via functional generalization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 10460–10469, 2020. 2

[2] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey 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 16*, pages 213–229. Springer, 2020. 1, 2

[3] Yu-Wei Chao, Yunfan Liu, Xieyang Liu, Huayi Zeng, and Jia Deng. Learning to detect human-object interactions. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*, 2018. 5

[4] Junwen Chen and Keiji Yanai. Qahoi: query-based anchors for human-object interaction detection. *arXiv preprint arXiv:2112.08647*, 2021. 2, 6

[5] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2

[6] Hao-Shu Fang, Jinkun Cao, Yu-Wing Tai, and Cewu Lu. Pairwise body-part attention for recognizing human-object interactions. In *Proceedings of the European conference on computer vision (ECCV)*, pages 51–67, 2018. 2

[7] Hao-Shu Fang, Yichen Xie, Dian Shao, and Cewu Lu. Dirv: Dense interaction region voting for end-to-end human-object interaction detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1291–1299, 2021. 2

[8] Chen Gao, Yuliang Zou, and Jia-Bin Huang. ican: Instance-centric attention network for human-object interaction detection. *arXiv preprint arXiv:1808.10437*, 2018. 2

[9] Chen Gao, Jiarui Xu, Yuliang Zou, and Jia-Bin 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 16*, pages 696–712. Springer, 2020. 2

[10] Rohit Girdhar and Deva Ramanan. Attentional pooling for action recognition. *Advances in neural information processing systems*, 30, 2017. 2

[11] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015. 2

[12] Saurabh Gupta and Jitendra Malik. Visual semantic role labeling. *arXiv preprint arXiv:1505.04474*, 2015. 5

[13] Tanmay Gupta, Alexander Schwing, and Derek Hoiem. No-frills human-object interaction detection: Factorization, layout encodings, and training techniques. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9677–9685, 2019. 2

[14] Zhi Hou, Baosheng Yu, Yu Qiao, Xiaojiang Peng, and Dacheng Tao. Detecting human-object interaction via fabricated compositional learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14646–14655, 2021. 6

[15] Ding Jia, Yuhui Yuan, Haodi He, Xiaopei Wu, Haojun Yu, Weihong Lin, Lei Sun, Chao Zhang, and Han Hu. Detrs with hybrid matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19702–19712, 2023. 1, 5

[16] Bumsoo Kim, Taeho Choi, Jaewoo Kang, and Hyunwoo 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 16*, pages 498–514. Springer, 2020. 2

[17] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR, 2022. 2

[18] Yong-Lu Li, Siyuan Zhou, Xijie Huang, Liang Xu, Ze Ma, Hao-Shu Fang, Yanfeng Wang, and Cewu Lu. Transferable interactiveness knowledge for human-object interaction detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3585–3594, 2019. 2

[19] Yong-Lu Li, Xinpeng Liu, Han Lu, Shiyi Wang, Junqi Liu, Jiefeng Li, and Cewu Lu. Detailed 2d-3d joint representation for human-object interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10166–10175, 2020. 2

[20] Yue Liao, Si Liu, Fei Wang, Yanjie Chen, Chen Qian, and Jiashi Feng. Ppdm: Parallel point detection and matching for real-time human-object interaction detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 482–490, 2020. 2

[21] Yue Liao, Aixi Zhang, Miao Lu, Yongliang Wang, Xiaobo Li, and Si Liu. Gen-vlkt: Simplify association and enhance interaction understanding for hoi detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20123–20132, 2022. 6

[22] JunYi Lim, Vishnu Monn Baskaran, Joanne Mun-Yee Lim, KokSheik Wong, John See, and Massimo Tistarelli. Ernet: An efficient and reliable human-object interaction detection network. *IEEE Transactions on Image Processing*, 32:964–979, 2023. 2

[23] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 5

[24] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 6

[25] Ye Liu, Junsong Yuan, and Chang Wen Chen. Consnet: Learning consistency graph for zero-shot human-object interaction detection. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 4235–4243, 2020. 2

[26] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 2

[27] Shuailei Ma, Yuefeng Wang, Shanze Wang, and Ying Wei. Fgahoi: Fine-grained anchors for human-object interaction detection. *arXiv preprint arXiv:2301.04019*, 2023. 2, 6

[28] Jeeseung Park, Jin-Woo Park, and Jong-Seok Lee. Viplo: Vision transformer based pose-conditioned self-loop graph for human-object interaction detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17152–17162, 2023. 2

[29] Siyuan Qi, Wenguan Wang, Baoxiong Jia, Jianbing Shen, and Song-Chun Zhu. Learning human-object interactions by graph parsing neural networks. In *Proceedings of the European conference on computer vision (ECCV)*, pages 401–417, 2018. 2

[30] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2

[31] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 2

[32] Masato Tamura, Hiroki Ohashi, and Tomoaki Yoshinaga. Qpic: Query-based pairwise human-object interaction detection with image-wide contextual information. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10410–10419, 2021. 2, 6

[33] Raphael Tang, Linqing Liu, Akshat Pandey, Zhiying Jiang, Gefei Yang, Karun Kumar, Pontus Stenetorp, Jimmy Lin, and Ferhan Ture. What the daam: Interpreting stable diffusion using cross attention. *arXiv preprint arXiv:2210.04885*, 2022. 2

[34] Bo Wan, Desen Zhou, Yongfei Liu, Rongjie Li, and Xuming He. Pose-aware multi-level feature network for human object interaction detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9469–9478, 2019. 2

[35] Hai Wang, Wei-shi Zheng, and Ling Yingbiao. 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 16*, pages 248–264. Springer, 2020. 2

[36] Tiancai Wang, Rao Muhammad Anwer, Muhammad Haris Khan, Fahad Shahbaz Khan, Yanwei Pang, Ling Shao, and Jorma Laaksonen. Deep contextual attention for human-object interaction detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5694–5702, 2019. 2

[37] Tiancai Wang, Tong Yang, Martin Danelljan, Fahad Shahbaz Khan, Xiangyu Zhang, and Jian Sun. Learning human-object interaction detection using interaction points. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4116–4125, 2020. 2

[38] Bingjie Xu, Yongkang Wong, Junnan Li, Qi Zhao, and Mohan S Kankanhalli. Learning to detect human-object interactions with knowledge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019. 2

[39] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. Vitpose: Simple vision transformer baselines for human pose estimation. *arXiv preprint arXiv:2204.12484*, 2022. 3

[40] Jie Yang, Bingliang Li, Fengyu Yang, Ailing Zeng, Lei Zhang, and Ruimao Zhang. Boosting human-object interaction detection with text-to-image diffusion model. *arXiv preprint arXiv:2305.12252*, 2023. 2, 3, 6

[41] Hangjie Yuan, Jianwen Jiang, Samuel Albanie, Tao Feng, Ziyuan Huang, Dong Ni, and Mingqian Tang. Rlip: Relational language-image pre-training for human-object interaction detection. *Advances in Neural Information Processing Systems*, 35:37416–37431, 2022. 2, 3

[42] Hangjie Yuan, Shiwei Zhang, Xiang Wang, Samuel Albanie, Yining Pan, Tao Feng, Jianwen Jiang, Dong Ni, Yingya Zhang, and Deli Zhao. Rlipv2: Fast scaling of relational language-image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21649–21661, 2023. 2, 3

[43] Aixi Zhang, Yue Liao, Si Liu, Miao Lu, Yongliang Wang, Chen Gao, and Xiaobo Li. Mining the benefits of two-stage and one-stage hoi detection. *arXiv preprint arXiv:2108.05077*, 2021. 2, 6

[44] Frederic Z Zhang, Dylan Campbell, and Stephen Gould. Spatially conditioned graphs for detecting human-object interactions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13319–13327, 2021. 2

[45] Frederic Z Zhang, Dylan Campbell, and Stephen Gould. Efficient two-stage detection of human-object interactions with a novel unary-pairwise transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20104–20112, 2022. 1, 2, 3, 4, 5, 6, 7

[46] Frederic Z Zhang, Yuhui Yuan, Dylan Campbell, Zhuoyao Zhong, and Stephen Gould. Exploring predicate visual context in detecting of human-object interactions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10411–10421, 2023. 1, 3, 6, 7, 8

[47] Yong Zhang, Yingwei Pan, Ting Yao, Rui Huang, Tao Mei, and Chang-Wen Chen. Exploring structure-aware transformer over interaction proposals for human-object interaction detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19548–19557, 2022. 6

[48] Xubin Zhong, Xian Qu, Changxing Ding, and Dacheng Tao. Glance and gaze: Inferring action-aware points for one-stage human-object interaction detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13234–13243, 2021. 2

[49] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable trans-

formers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020. 1

[50] Cheng Zou, Bohan Wang, Yue Hu, Junqi Liu, Qian Wu, Yu Zhao, Boxun Li, Chenguang Zhang, Chi Zhang, Yichen Wei, and Jian Sun. End-to-end human object interaction detection with hoi transformer. In *CVPR*, 2021. 2

[51] Cheng Zou, Bohan Wang, Yue Hu, Junqi Liu, Qian Wu, Yu Zhao, Boxun Li, Chenguang Zhang, Chi Zhang, Yichen Wei, et al. End-to-end human object interaction detection with hoi transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11825–11834, 2021. 6