

HORP: Human-Object Relation Priors Guided HOI Detection

Pei Geng, Jian Yang, Shanshan Zhang*

PCA Lab, Key Lab of Intelligent Perception and Systems for High-Dimensional Information
 of Ministry of Education, and Jiangsu Key Lab of Image and Video Understanding for Social Security,
 School of Computer Science and Engineering, Nanjing University of Science and Technology

{peigeng, csjyang, shanshan.zhang@njust.edu.cn}

Abstract

Human-Object Interaction (HOI) detection aims to predict the $\langle \text{Human}, \text{Interaction}, \text{Object} \rangle$ triplets, where the core challenge lies in recognizing the interaction of each human-object pair. Despite recent progress thanks to more advanced model architectures, HOI performance remains unsatisfactory. In this work, we first perform some failure analysis and find that the accuracy of the no-interaction category is extremely low, largely hindering the improvement of overall performance. We further look into the error types and find the mis-classification between no-interaction and with-interaction ones can be handled by human-object relation priors. Specifically, to better distinguish no-interaction from direct interactions, we propose 3D location prior, which indicates the distance between human and object; as of no-interaction vs. indirect interactions, we propose gaze area prior, which denotes whether human can see the object or not. The above two types of human-object relation priors are represented by text and are combined with the original visual features, generating multi-modal cues for interaction recognition. Experimental results on the HICO-DET and V-COCO datasets demonstrate that our proposed human-object relation priors are effective and our method HORP surpasses previous methods under various settings and scenarios. In particular, the usage of our priors significantly enhances the model's recognition ability for the no-interaction category. Code is available at <https://github.com/namegp/HORP>.

1. Introduction

Human-Object Interaction Detection (HOI) has received widespread attention as an essential computer vision task in recent years. Based on object detection, HOI adds more advanced action understanding and is widely applied in tasks such as image retrieval [8, 13], image captioning [3, 26, 41],

*Corresponding author.

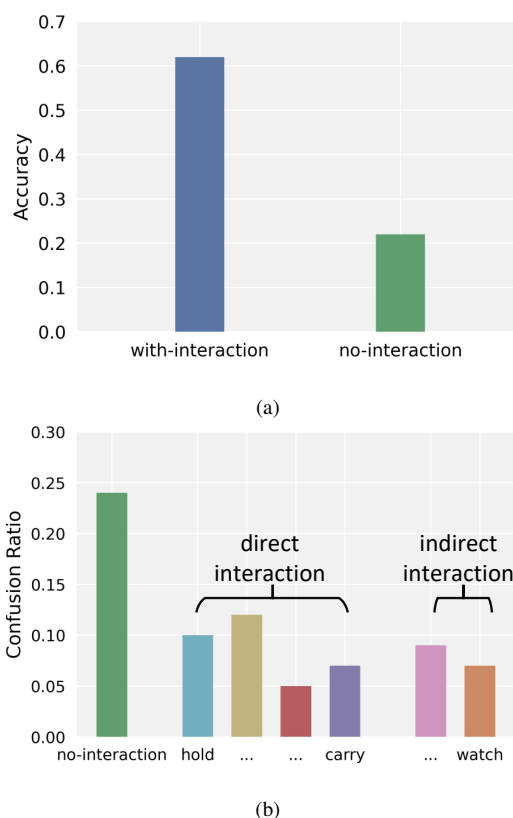


Figure 1. The accuracy of no-interaction is much lower than that of with-interaction (a); those no-interaction samples are easily mis-classified as both direct and indirect interactions (b).

and human action recognition [16, 20, 25, 31, 49]. The core of the HOI task is to predict the $\langle \text{Human}, \text{Interaction}, \text{Object} \rangle$ triplets in the image, which consists of two key subtasks: detecting humans and objects [18, 19, 47, 48]; categorizing the interaction of each human-object pair, which is vital for the HOI task due to the diversity of interaction categories and the complexity of the scenes.

Existing methods [5, 15, 37, 51, 53] usually improve HOI detection through more accurate interaction recognition. Despite significant advancements in previous work,

the current performance remains unsatisfactory and does not meet the requirements of real-world applications. To locate the bottleneck, we analyze the accuracy across different interaction categories. As shown in Figure 1(a), the accuracy of the no-interaction category is significantly lower (~ 40 pp) than that with-interaction categories, hindering the improvement of the overall performance.

We try to understand the challenges of no-interaction by observing its error types. As shown in Figure 1(b), the no-interaction category is usually mistakenly recognized as those with-interaction ones, which can be divided into direct-interaction (e.g., carry, hold) and indirect-interaction (e.g., watch, inspect) groups based on whether there exists direct body contact. We further look into some examples from the above two groups of interactions by comparing them with the no-interaction ones (Figure 2), and obtain the following interesting findings. (1) For direct-interactions (e.g., 'carry'), the human and the object must have physical contact and thus are close to each other in the 3D space (Figure 2a(bottom)); in contrast, for no-interaction, the person stays far away from the object (Figure 2a(top)). (2) For indirect-interactions (e.g., watch), the human and the object have no physical contact but the object is within the human's gaze area (Figure 2b(bottom)); in contrast, for no-interaction, the object stays out of the human's gaze area (Figure 2b(top)). The above findings strongly motivate us to use the 3D location prior to distinguish direct-interaction with no-interaction, and human gaze area prior to separate indirect-interaction from no-interaction. Therefore, in this work we explore how to incorporate the two types of priors into HOI detection models to better distinguish different types of interactions.

In this work, we propose a Human-Object Relation Prior (HORP) guided approach that aims to enhance HOI detection by improving the model's ability to distinguish different types of interactions. Specifically, we design two types of human-object relation priors, the human-object 3D location prior and the human gaze area prior. Considering the 2D location on the image plane is sometimes misleading (objects close to each other may appear at different depths), we shift to the 3D space, so as to describe the locations of objects in the real world more accurately. The 3D location is represented by a combination of 2D spatial location on the image plane, and depth, both in a relative manner between the human and object. On the 2D image plane, we use the IoU value to describe their relative 2D location; considering it is too expensive to measure the depth values by an external model, we roughly estimate the relative depth by 'person + number of pixels' and 'object category + number of pixels'. One can imagine that a person with 200×100 pixels is unable to touch a car with 50×100 pixels as they vary too much in depth. For the gaze prior, we introduce a gaze detector to predict the gaze area of the human, which

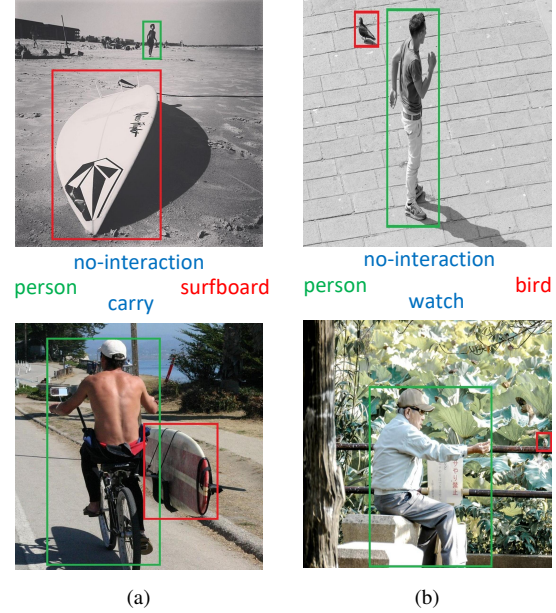


Figure 2. Comparison of different types of interactions under the same human and object pair. (a) no-interaction vs. direct interaction; (b) no-interaction vs. indirect interaction.

can be used to estimate whether the object is within the human's gaze area. Then, we transform the two kinds of human-object relation priors to text respectively, and pass the text to a frozen text encoder. After that, we fuse the vanilla visual features and the text features from two priors of each human-object pair and treat them as queries in the interaction decoding phase, where the interaction category is predicted. In this way, we incorporate two types of human-object relation priors into HOI detection to improve the overall performance.

In summary, our contributions are as follows:

- We for the first time point out the mis-classification of no-interaction category is an overlooked problem in HOI detection, which hinders the overall performance improvement by a large extent. We further look into its error types and find this problem can be addressed by introducing prior knowledge in terms of 3D location and human gaze.
- Based on the above findings, we propose a Human-Object Relation Prior (HORP) guided approach that integrates the Human-Object 3D Location Prior and Human Gaze Area Prior into the HOI detection model.
- Experiments on the HICO-DET and V-COCO datasets demonstrate that our method significantly enhances the recognition of no-interaction category and outperforms state-of-the-art methods in overall performance.

2. Related Work

In this section, we present methods related to this work, covering methods based on the two-stage paradigm and methods that utilize visual-language models.

2.1. Visual HOI Detection

Existing HOI detection methods can be categorized into one-stage and two-stage paradigms. One-stage approaches [5, 14, 15, 52] perform the detection of human-object pairs and the recognition of interaction actions in parallel through an end-to-end manner, which is architecturally simple and more efficient. For example, DIRV [5] captures visual features at various scales of each human-object pair and utilizes the overlapped interaction region instead of NMS to compensate for the detection deficiency. HOTR [14] is the first Transformer-based method that links instance and interaction decoders using HO pointers to predict HOI triples.

Though one-stage approaches are structurally simple and fast, they are not easily scalable and accuracy is not satisfactory. In contrast, two-stage detection approaches [10, 35, 37, 51, 53] first perform object detection using an off-the-shelf detector, and then external knowledge or additional features can be introduced to facilitate the classification of interactions. Recent researches [45, 46, 50] prove that the separation of object detection and interaction inference can improve the accuracy and flexibility of HOI tasks. Specifically, UPT [45] refines DETR outputs with an extra Transformer layer and integrates pairwise bounding box encoding to jointly infer appearance and spatial relations. STIP [50] splits HOI prediction into two phases: generating interaction proposals and converting them into HOI predictions using a structure-aware Transformer. The two-stage approaches are easily extensible and allow the fusion of other features (appearance features, spatial features for detecting bounding boxes, text embedding, etc.) at the second stage to predict the interaction of each pair. Therefore, our approach is based on the two-stage paradigm.

2.2. VLM-Based HOI Detection

The Visual-Language models (VLMs) [12, 29] have demonstrated strong capability in capturing high-quality contextual features, making language supervision effective and transferable to downstream tasks. Many current HOI methods [17, 22, 27, 28, 34, 36] attempt to leverage the knowledge of VLM for interaction recognition in HOI tasks. For example, GEN-VLKT [22] uses knowledge transfer from CLIP to enhance interaction understanding. ADA-CM [17] introduces a concept-guided memory combining visual and semantic knowledge to boost HOI performance. BCOM [36] relies solely on CLIP’s visual encoder for spatial feature extraction and designs a module for adaptive occlusion handling. ViPLIO [27] replaces traditional CNNs with CLIP’s image encoder and introduces a pose-based graph for interaction recognition. DOQ [28] uses oracle queries with word embeddings from the text encoder for decoder initialization. AGER [34] dynamically clusters patch tokens and aligns cluster centers using text guidance to extract instance-level features. In contrast, we use VLMs for a dif-

ferent purpose. Given that text embeddings offer more flexibility, enabling fine-grained semantic descriptions and better capturing human-object relational information. Therefore, we incorporate textual priors into well-designed textual templates and combine them with the visual features of human-object pairs to guide the learning of the interaction decoder.

3. Methodology

Problem Definition. Given an input image X , the model needs to predict a set of HOI instances $T = (b^h, b^o, c^o, c^{h,o})$, where b^h , b^o , c^o and $c^{h,o}$ denote the human bounding box, the object bounding box, the object category, and the interaction category of one human-object pair, respectively.

3.1. Overview

We adopt a two-stage detection paradigm considering its superior performance and scalable intermediate features. The overall architecture of our method is shown in Figure 3.

At the first stage, for an image $X \in \mathbb{R}^{H \times W \times C}$, we use a backbone network to extract the visual features $F_v \in \mathbb{R}^{H_0 \times W_0 \times C_0}$, and then use stacked Transformer encoder and decoder to localize the humans $B^h \in \mathbb{R}^{N_h \times 4}$ and objects $\{B^o \in \mathbb{R}^{N_o \times 4}, C^o \in \mathbb{R}^{N_o \times 1}\}$, where N_h and N_o denote the number of humans and objects, respectively.

At the second stage, we perform interaction recognition for each human-object pair using two kinds of priors. Specifically, for 3D location prior, we use the IoU $\mathcal{I}_{h,o}$ of the human-object pair to represent 2D spatial location on the image plane; then, we use a combination of human/object category c^o and their scale values $\mathcal{S}_h/\mathcal{S}_o$ to represent the relative depth between human and object; the above information is filled into a designed text template and encoded to get the 3D location prior features $\mathcal{Z}_{loc}$. In addition, we construct the gaze area prior by calculating the proportion $\mathcal{G}_{pro}$ of the object in the human gaze area; then, we fill the designed text template and encode the text to get the gaze area prior features $\mathcal{Z}_{gaze}$. After that, we follow the previous approaches [45, 46] and employ a Transformer-based Decoder for interaction recognition. To facilitate more accurate interaction decoding, we construct a multi-modal query by concatenating both human-object relation prior features with the visual features of the human-object pairs. Finally, we use a feedforward network (FFN) to predict the interaction category to derive the HOI triplet.

3.2. Human-Object 3D Location Prior

To better distinguish no-interaction from direct interactions, we propose to use the human-object 3D location prior for assistance, as it is common sense that the person cannot interact directly with an object that is too far away. Concretely, simply using the 2D spatial location of human-

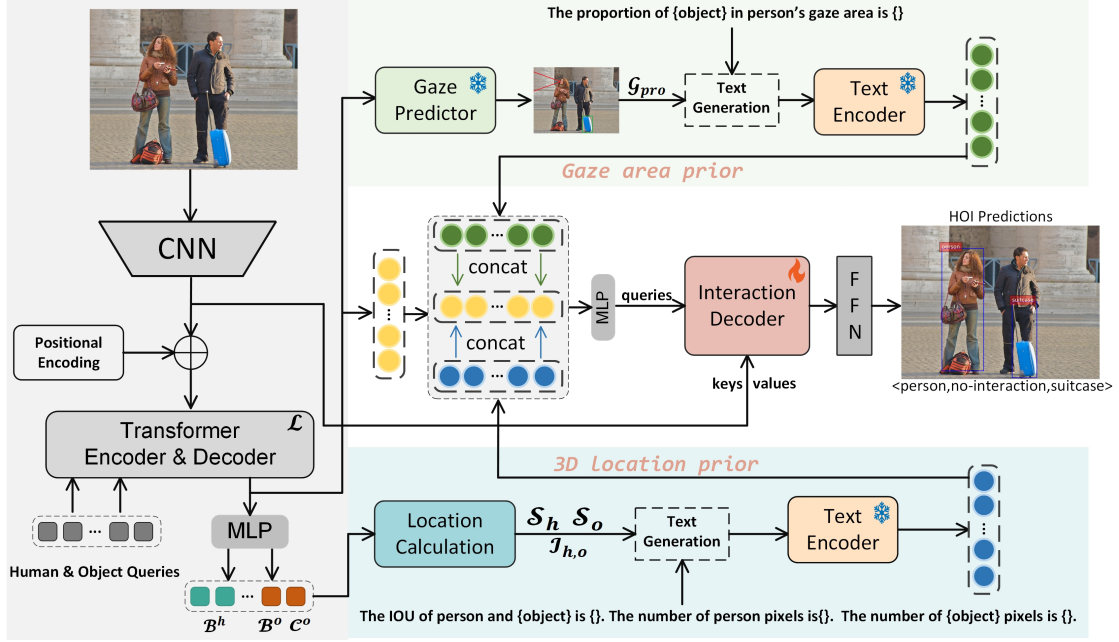


Figure 3. The overall architecture of HORP. The key components include the human gaze area prior and the human-object 3D location prior generation. Both two priors are then integrated with visual features in the form of text embeddings and used as queries in the interaction decoding stage for interaction recognition.

object pairs is not sufficient, due to the fact that the human-object pairs close on the image plane may stay far away in the real 3D space, see Figure 4 for an illustration. Therefore, we use a combination of 2D spatial location on the image plane along with estimated depth to represent 3D location of each human-object pair.

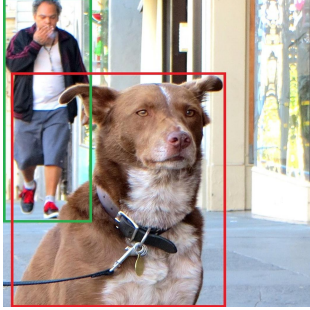


Figure 4. An example of the no-interaction category: a person and a dog are with a high IoU, but are far away from each other.

First, we calculate the IOU $\mathcal{I}_{h,o}$ between human and object bounding boxes to represent their relative 2D spatial location. Then, the scale values $\mathcal{S}_h$ and $\mathcal{S}_o$, are computed by normalizing the number of pixels inside each box w.r.t. the image size as follows:

$$\mathcal{S}_h = \frac{\mathcal{N}^h}{\mathcal{N}_{\text{img}}}, \mathcal{S}_o = \frac{\mathcal{N}^o}{\mathcal{N}_{\text{img}}}, \quad (1)$$

where $\mathcal{N}^h$, $\mathcal{N}^o$ and $\mathcal{N}_{\text{img}}$ denote the number of pixels within the bounding box of the human, object, and the total number of pixels, respectively. And the scale values

are combined with categories to represent relative depth between the human and object.

Finally, we fill the spatial location and depth information into the designed template $\mathcal{T}_{loc}$: “The IoU of person and $\{C^o\}$ is $\{\mathcal{I}_{h,o}\}$. The number of person pixels is $\{\mathcal{S}_h\}$. The number of $\{C^o\}$ pixels is $\{\mathcal{S}_o\}$.”, and send the text to the encoder to obtain the human-object 3D location prior $\mathcal{Z}_{loc}$ as follows:

$$\mathcal{Z}_{loc} = \text{TextEnc}(\text{Embed}(\mathcal{T}_{loc})). \quad (2)$$

3.3. Human Gaze Area Prior

Since indirect interactions like inspect and watch, can only happen when the object is within the human’s gaze area, we propose to use human gaze area prior to better distinguish no-interaction from indirect interactions.

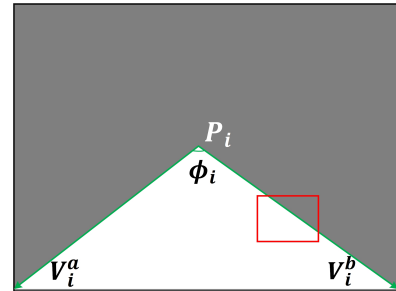


Figure 5. An illustration of the human gaze area, where the red box denotes the object bounding box and the white area denotes the gaze area.

First, we employ an off-the-shelf gaze cone predictor [33] to obtain a gaze cone with a pre-defined angle of ϕ_i , representing the gaze area based on the visual features of each human-object pair. The gaze cone is noted as $a_i^h = (P_i, V_i^a, V_i^b)$, where P_i denotes the vertex of the gaze cone; V_i^a and V_i^b denote the two vectors along the boundaries of the gaze cone. In Figure 5 we show an illustration, where the white area corresponds to the human gaze area.

After that, we further compute how much of the object can be seen by human. We devise an efficient way for computation: we binarize the gaze map as shown in Figure 5, and then count how many white pixels fall within the object box. The proportion $\mathcal{G}_{pro}$ can be computed as follows:

$$\mathcal{G}_{pro} = \frac{\mathcal{N}_{white}}{\mathcal{N}^o}, \quad (3)$$

where $\mathcal{N}_{white}$ and $\mathcal{N}^o$ indicate the number of white pixels fall within the object box, the size of the object box w.r.t. number of pixels. Considering efficiency, we downsample the original gaze map by $100\times$ before computation.

Finally, we fill in the template $\mathcal{T}_{gaze}$ ‘‘The proportion of $\{C^o\}$ in person’s gaze area is $\{\mathcal{G}_{pro}\}$ ’’ and obtain the human gaze area prior $\mathcal{Z}_{gaze}$ in the form of text encoding:

$$\mathcal{Z}_{gaze} = TextEnc(Embed(\mathcal{T}_{gaze})). \quad (4)$$

3.4. Multi-Modal Query Construction

Previous approaches usually initialize the query by random Gaussian noise, or visual features and gradually aggregate useful information during the training process. However, this manner may fail to capture critical information especially under complex scenarios. Aiming for more effective learning and recognition, we construct more informative queries by combining the visual features and the two kinds of human-object relation priors proposed in Section 3.2 and Section 3.3.

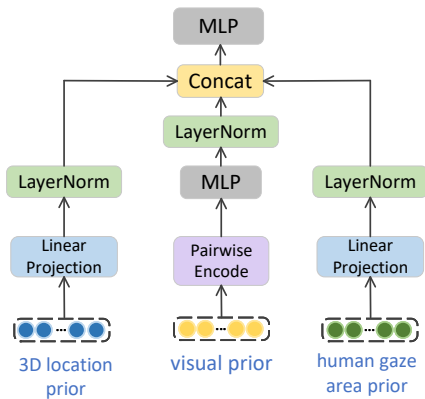


Figure 6. Construction of multi-modal queries at the interaction decoding stage.

As shown in Figure 6, we first pairwise encode the visual features F_v of the human-object pairs, and adjust the feature

dimensions with an MLP, to normalize the visual features. For 3D location prior $\mathcal{Z}_{loc}$ and gaze area prior $\mathcal{Z}_{gaze}$, we adopt a linear projection layer $Proj(\cdot)$ to map the text prior into the visual feature space and perform normalization:

$$\mathcal{Z}'_{gaze} = Proj(\mathcal{Z}_{gaze}), \mathcal{Z}'_{loc} = Proj(\mathcal{Z}_{loc}). \quad (5)$$

Then, we concatenate the visual features with the textual prior features and use them to generate queries via an MLP, where the visual features are used as the keys and values.

3.5. Optimization Objective for Interaction Recognition

After the interaction decoder, we use an MLP as the classifier for interaction recognition. To alleviate imbalance among categories, we use Focal Loss [30] at the second stage as the optimization objective, which can be written as follows:

$$L_{cls} = \frac{1}{\sum_{i=1}^N \sum_{c=1}^C y_i^c} \sum_{i=1}^N \sum_{c=1}^C FL(\hat{y}_i^c, y_i^c), \quad (6)$$

where C is the number of interaction categories; N is the number of candidate human-object pairs; y_i^c means that the ground truth of the i -th human-object pair includes the c -th interaction category; $\hat{y}_i^c$ is the corresponding probability distribution from the classifier.

4. Experiments

In this section, we first introduce the datasets (Section 4.1) and experimental details (Section 4.2). Next, we conduct ablation studies (Section 4.3). After that, we compare the proposed HORP with state-of-the-art methods (Section 4.4). Finally, we present qualitative results (Section 4.5).

4.1. Datasets and Metrics

Datasets. We fully evaluate our method on two public datasets HICO-DET [2] and V-COCO [9]. HICO-DET comprises 47,776 images, with 38,118 designated for training and 9,658 for testing. It includes 600 HOI categories formed by 117 interaction classes and 80 object categories, which align with those on the MS-COCO dataset. The dataset provides over 150,000 annotated human-object pairs, offering a rich resource for evaluating various interactions. Furthermore, V-COCO, a subset of MS-COCO, contains 10,346 images divided into 2,533 for training, 2,867 for validation, and 4,946 for testing. This dataset features the same 80 object categories as HICO-DET, but with only 29 interaction categories, making it a smaller yet significant resource for HOI studies.

Evaluation Metrics. For evaluation, we adhere to established protocols, reporting mean Average Precision (mAP) as the primary metric. A prediction of a HOI triplet is classified as a true positive when both the predicted human and

object bounding boxes achieve an Intersection over Union (IoU) greater than 0.5 against the ground truth, alongside accurate classification of the interaction. On HICO-DET, we evaluate performance under two settings: the Default setting, which assesses predictions across the entire test set, and the Known Object setting, which focuses on images containing specific target object categories. We report mAP across three subsets: all 600 HOI categories (Full), 138 categories with fewer than 10 training samples (Rare), and the remaining 462 categories (Non-Rare). For V-COCO, we present mAP for two scenarios: one that includes predictions for occluded object bounding boxes and another that disregards these predictions.

4.2. Implementation Details

For a fair comparison with state-of-the-art methods, we employ the object detector DETR pre-trained on MS-COCO and all learnable parameters in DETR are frozen during training. For the gaze predictor, we use the off-the-shelf gaze cone predictor from the gaze target detection model [33] and fix the parameters. We employ the pre-trained ViT-32/B CLIP to generate text embeddings and keep its parameters fixed. After detection, we filter out instances with confidence scores less than 0.2 and sample a minimum of 3 and a maximum of 15 human and object instances, respectively. We apply the data augmentation techniques used in previous works [45, 54] to perform random cropping, size rescaling, and random color jittering on the input images. We use AdamW [23] with an initial learning rate and a weight decay of 1×10^{-4} as the optimizer for the model, and employ a multi-step learning rate decay mechanism to drop the learning rate by a factor of 5 at the 20^{th} epoch. All experiments are conducted with a batch size of 16 on 4 GeForce 3090 GPUs.

Table 1. Comparison of performance using different human-object relation priors under the HICO-DET Default Setting.

#	relational priors		Default Setting		
	gaze	location	Full	Rare	Non-rare
A1			35.10	33.42	35.60
A2	✓		36.85 (+1.75)	34.73	37.48
A3		✓	37.24 (+2.14)	35.09	37.88
A4	✓	✓	38.61 (+3.51)	36.14	39.34

4.3. Ablation Study

In this subsection, we conduct ablation experiments to analyze the effectiveness of our proposed method. For simplicity, we adopt ResNet-50 as the backbone, and all ablation studies are performed on the HICO-DET and V-COCO datasets.

Impact of Different Priors. To assess the impact of the proposed two types of human-object relation priors, we add

Table 2. Comparison of accuracy for the no-interaction category using different human-object relation priors under the HICO-DET Default Full Setting.

#	relational priors		Default Setting
	gaze	location	Full
A1			16.32
A2	✓		19.65 (+3.33)
A3		✓	20.16 (+3.84)
A4	✓	✓	23.11 (+6.79)

Table 3. Ablation study of the 3D location prior under the HICO-DET Default Setting.

#	location prior		Default Setting		
	spatial	depth	Full	Rare	Non-rare
B1			36.85	34.73	37.48
B2	✓		37.22 (+0.37)	35.34	37.78
B3		✓	37.45 (+0.60)	35.60	38.01
B4	✓	✓	38.61 (+1.76)	36.14	39.34

them one by one on top of the baseline detector. The results are presented in Table 1 and Table 2, and we have the following observations: (1) when we use the 3D location prior only, the performance improves by 2.14 pp w.r.t. mAP on Full and by 3.84 pp w.r.t. AP for the no-interaction category, (2) when we use the gaze area prior only, the performance improves by 1.75 pp w.r.t. mAP on Full and by 3.33 pp w.r.t. AP for the no-interaction category, (3) when both priors are used, we achieve 38.61% mAP on Full and 23.11% mAP for no-interaction, outperforming the baseline by 3.51 pp and 6.79 pp, respectively. These results indicate that each prior is helpful for interaction recognition, especially for the more challenging no-interaction category.

Design Choices of 3D Location Prior. To evaluate the necessity of using both 2D spatial and depth information to represent 3D location prior, we perform the following ablation study. As shown in Table 3, when we use solely 2D spatial prior or depth prior, the improvement is rather small (0.37 pp from 2D spatial and 0.60 pp from depth), indicating that either of them is not good enough to represent 3D location. When we use both of them, the improvement becomes more notable, i.e. 1.76 pp w.r.t. mAP on Full. These results indicate that it is necessary to use a combination of 2D spatial and depth to represent the relative locations of human and object.

Design Choices of Gaze Area Prior. In this work we explicitly represent how much of the object appears in the human gaze area with text; alternatively, a more naive way would be simply feeding the gaze map to the network, allowing it to interpret the information. We compare the above two ways of using gaze area prior by experiments. As shown in Figure 7, we find that the explicit text prior is more effective than the implicit visual prior and the gap is as large as 3 pp across different settings. These results indicate that

Table 4. Comparison on the HICO-DET and V-COCO datasets under various settings and scenarios. The best results are highlighted in **bold**, and the second-best results are underlined.

Method	Feature Extractor	HICO-DET						V-COCO	
		Default Setting			Known-Object Setting			S_1	S_2
		Full	Rare	Non-rare	Full	Rare	Non-rare		
InteractNet [7]	ResNet-50-FPN	9.94	7.16	10.77	-	-	-	40.0	-
HOTR [14]	ResNet-50	25.10	17.34	27.42	-	-	-	55.2	64.4
RLIP [42]	ResNet-50	32.84	26.85	34.63	-	-	-	61.9	64.2
PVIC [46]	ResNet-50	34.69	32.14	35.45	38.14	35.38	38.97	62.8	67.8
CDN [44]	ResNet-50	31.78	27.55	33.05	34.53	29.73	35.96	62.3	64.4
STIP [50]	ResNet-50	32.22	28.15	33.43	35.29	31.43	36.45	66.0	70.7
CATN [4]	ResNet-50	31.86	25.15	33.84	34.44	27.69	36.45	60.1	-
IDN [21]	ResNet-50	24.58	20.33	25.86	27.89	23.64	29.16	53.3	60.3
AGER [34]	ResNet-50	36.75	33.53	37.71	39.84	35.58	40.23	65.7	69.7
ViPLO [27]	ResNet-50+CLIP	37.22	35.45	37.75	40.61	38.82	41.15	62.2	68.0
GEN-VLKT [22]	ResNet-50+CLIP	33.75	29.25	35.10	36.78	32.75	37.99	62.4	64.5
ADA-CM [17]	ResNet-50+CLIP	38.40	37.52	38.66	-	-	-	58.6	64.0
SICHOI [24]	ResNet-50+BLIP+GPT	41.79	42.38	41.61	44.27	43.64	44.46	67.9	72.8
QPIC [32]	ResNet-101	29.90	23.92	31.69	32.38	26.06	34.27	58.8	61.0
UPT [45]	ResNet-101	32.62	28.62	33.81	36.08	31.41	37.47	61.3	67.1
HOI-Trans [54]	ResNet-101	26.61	19.15	28.84	29.13	20.98	31.57	52.9	-
CQL [39]	ResNet-101	36.03	33.16	36.89	38.82	35.51	39.81	66.5	69.9
RmLR [1]	ResNet-101	37.41	28.81	39.97	38.69	31.27	40.91	64.2	70.2
RLIPv2 [43]	Swin-T	33.66	40.07	38.60	-	-	-	<u>68.8</u>	70.8
DiffHOI [40]	Swin-L	41.50	39.96	41.96	43.62	41.41	44.28	65.7	68.2
PVIC [46]	Swin-L	44.32	44.61	44.24	47.81	48.38	47.64	64.1	70.2
PAFR [38]	Swin-L	<u>46.01</u>	<u>46.74</u>	<u>45.80</u>	<u>49.50</u>	<u>50.59</u>	<u>49.18</u>	63.0	68.7
HORP (Ours)	ResNet-50+CLIP	38.61	36.14	39.34	40.98	38.25	41.79	65.6	68.3
HORP (Ours)	Swin-L+CLIP	47.53	46.81	47.74	51.24	50.78	51.38	68.9	<u>71.1</u>

the explicit gaze area prior in the text form provides more informative cues to the model, and thus enhances its ability to distinguish no-interaction from indirect interactions.

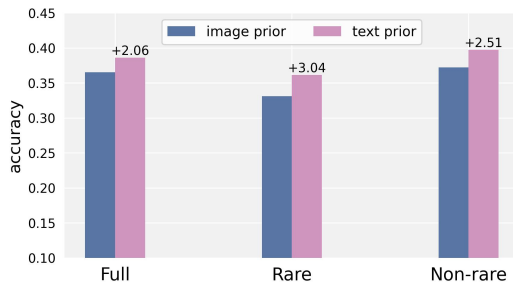


Figure 7. Comparison of using the gaze area prior in the explicit text form and in the implicit image form. Numbers are accuracy under the HICO-DET Default Setting.

Different Ways of Query Construction. To investigate the impact of different query construction methods on performance, we conduct experiments using four different variants and compare their results in Table 5. It can be observed that the performance stays stable under different query construction methods, with less than 1 pp variance. Such small

Table 5. Ablation study of different query construction methods under the HICO-DET Default Setting.

Method	Default Setting		
	Full	Rare	Non-rare
weighted sum	37.75	35.23	38.50
element-wise multiplication	36.86	34.79	37.48
average pooling	37.24	34.97	37.92
concatenation	38.61	36.14	39.34

gaps indicate that our proposed priors are not sensitive to the integration strategy. Among the four different ways, ‘concatenation’ performs the best as it preserves more information from different cues, so we use it as the default setting in our experiments.

4.4. Comparison with State-of-the-Art Methods

In this subsection, we report our experimental results on the HICO-DET and V-COCO datasets, and compare with state-of-the-art methods in Table 4.

On the HICO-DET dataset, our HORP method using ResNet-50 as backbone outperforms previous methods with the same backbone across all evaluation settings, except for

SICHOI [24], which employs BLIP’s image and text encoders and generates textual cues with GPT. Since some recent methods, e.g., RLIPv2, DiffHOI, PVIC, PAFR, use ViT as backbone, we also implement our HORP with Swin-L and compare with them. We can see our method surpasses the previously top method PAFR by 1.53 pp under the default full setting. On the V-COCO dataset, our HORP with Swin-L outperforms all previous methods under the more challenging scenario 1, and performs the second best under scenario 2. Overall, our HORP method achieves state-of-the-art results on two standard benchmarks, demonstrating the effectiveness of our proposed human-object relation priors.

4.5. Qualitative Analysis

We present several qualitative results on the challenging no-interaction category from the HICO-DET test set in Figure 8. These results demonstrate a significant improvement of our method in recognizing the no-interaction category. For instance, in Figure 8(a), the proposed HORP can infer that the vase is outside the person’s gaze region using the human gaze area prior, which increases the likelihood of predicting no-interaction; in Figure 8(b), the person and surfboard are not overlapping in 2D spatial location and the surfboard is not in the person’s gaze region, the model tends to predict no-interaction; in Figure 8(c), the cake and the person obviously vary a lot in terms of depth and thus no interaction could happen between them, in which case our designed 3D location prior is helpful; similarly, in Figure 8(d), there is a clear discrepancy in depth between the stop_sign and the person. In summary, our HORP effectively recognizes the no-interaction pattern between humans and objects.

5. Conclusion

In this paper, we propose a Human-Object Relation Prior Guided (HORP) approach to enhance the overall performance of HOI detection by improving the model’s ability to distinguish the no-interaction category. We design a Human-Object 3D Location Prior to differentiate between no-interaction and direct-interaction categories, and a Human Gaze Area Prior to distinguish between no-interaction and indirect-interaction categories. To better capture interaction cues, we encode the above two types of human-object relation priors into text features and fuse them with the original visual features, generating multi-modal cues to guide interaction recognition. Experimental results on the HICO-DET and V-COCO datasets show that our method significantly improves the recognition of the no-interaction category and enhances the overall performance.

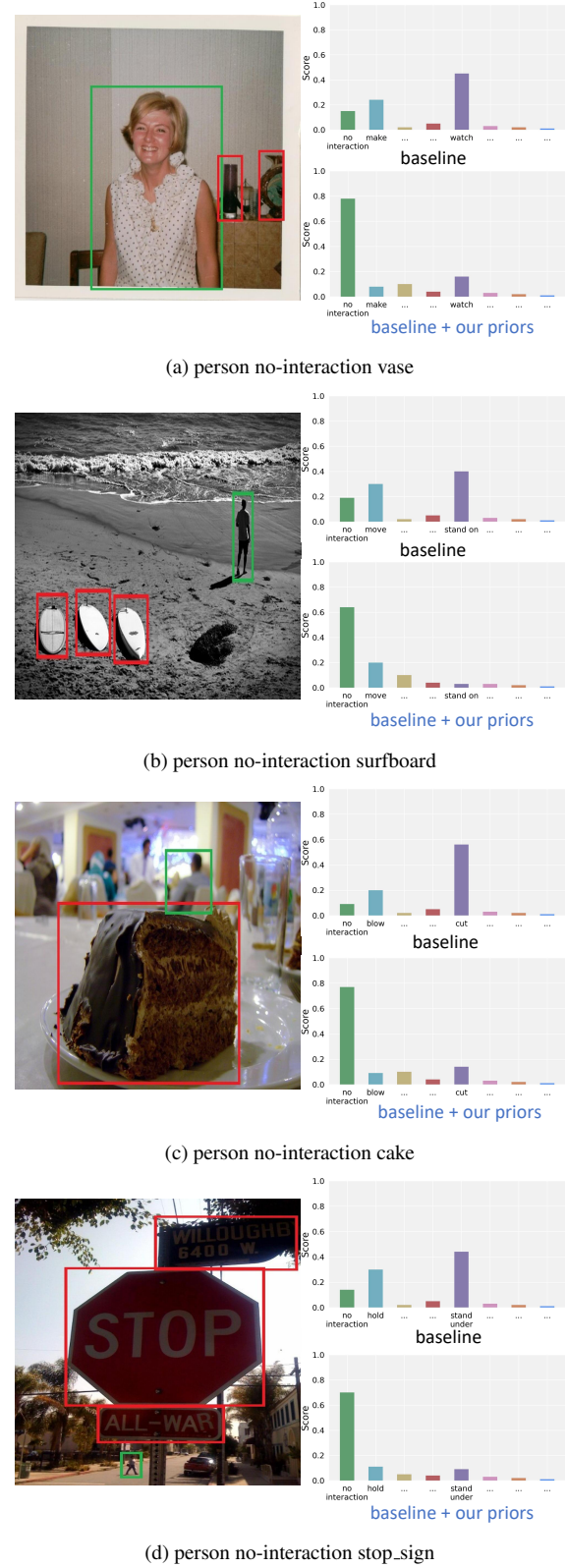


Figure 8. Qualitative results on the HICO-DET test set. Our method significantly improves the accuracy of the no-interaction category after integrating human-object relational priors.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (Grant No. 62322602, 62172225), and Natural Science Foundation of Jiangsu Province, China (Grant No. BK20230033).

References

- [1] Yichao Cao, Qingfei Tang, Feng Yang, Xiu Su, Shan You, Xiaobo Lu, and Chang Xu. Re-mine, learn and reason: Exploring the cross-modal semantic correlations for language-guided hoi detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. 7
- [2] Yu-Wei Chao, Yunfan Liu, Xieyang Liu, Huayi Zeng, and Jia Deng. Learning to detect human-object interactions. In *Ieee Winter Conference on Applications of Computer Vision*, 2018. 5
- [3] Shizhe Chen, Qin Jin, Peng Wang, and Qi Wu. Say as you wish: Fine-grained control of image caption generation with abstract scene graphs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020. 1
- [4] Leizhen Dong, Zhimin Li, Kunlun Xu, Zhijun Zhang, Luxin Yan, Sheng Zhong, and Xu Zou. Category-aware transformer network for better human-object interaction detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 7
- [5] 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*, 2021. 1, 3
- [6] Chen Gao, Jiarui Xu, Yuliang Zou, and Jia-Bin Huang. Drg: Dual relation graph for human-object interaction detection. In *Proceedings of the European Conference on Computer Vision*, 2020.
- [7] Georgia Gkioxari, Ross Girshick, Piotr Dollár, and Kaiming He. Detecting and recognizing human-object interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018. 7
- [8] Albert Gordo and Diane Larlus. Beyond instance-level image retrieval: Leveraging captions to learn a global visual representation for semantic retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2017. 1
- [9] Saurabh Gupta and Jitendra Malik. Visual semantic role labeling. *arXiv preprint arXiv:1505.04474*, 2015. 5
- [10] 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*, 2019. 3
- [11] Zhi Hou, Xiaojiang Peng, Yu Qiao, and Dacheng Tao. Visual compositional learning for human-object interaction detection. In *Proceedings of the European Conference on Computer Vision*, 2020.
- [12] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, 2021. 3
- [13] Justin Johnson, Ranjay Krishna, Michael Stark, Li-Jia Li, David Shamma, Michael Bernstein, and Li Fei-Fei. Image retrieval using scene graphs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2015. 1
- [14] Bumsoo Kim, Junhyun Lee, Jaewoo Kang, Eun-Sol Kim, and Hyunwoo J Kim. Hotr: End-to-end human-object interaction detection with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021. 3, 7
- [15] Sanghyun Kim, Deunsol Jung, and Minsu Cho. Relational context learning for human-object interaction detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 1, 3
- [16] Yu Kong and Yun Fu. Human action recognition and prediction: A survey. *International Journal of Computer Vision*, 130(5):1366–1401, 2022. 1
- [17] Ting Lei, Fabian Caba, Qingchao Chen, Hailin Jin, Yuxin Peng, and Yang Liu. Efficient adaptive human-object interaction detection with concept-guided memory. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. 3, 7
- [18] Gang Li, Xiang Li, Yujie Wang, Yichao Wu, Ding Liang, and Shanshan Zhang. Pseco: Pseudo labeling and consistency training for semi-supervised object detection. In *Proceedings of the European Conference on Computer Vision*, 2022. 1
- [19] Gang Li, Xiang Li, Yujie Wang, Shanshan Zhang, Yichao Wu, and Ding Liang. Knowledge distillation for object detection via rank mimicking and prediction-guided feature imitation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1306–1313, 2022. 1
- [20] Gang Li, Xiang Li, Shanshan Zhang, and Jian Yang. Towards more reliable evaluation in pedestrian detection by rethinking “ignore regions”. *Visual Intelligence*, 2(1):4, 2024. 1
- [21] Yong-Lu Li, Xinpeng Liu, Xiaoqian Wu, Yizhuo Li, and Cewu Lu. Hoi analysis: Integrating and decomposing human-object interaction. In *Advances in Neural Information Processing Systems*, 2020. 7
- [22] 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*, 2022. 3, 7
- [23] I Loshchilov. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6
- [24] Jinguo Luo, Weihong Ren, Weibo Jiang, Xi’ ai Chen, Qiang Wang, Zhi Han, and Honghai Liu. Discovering syntactic interaction clues for human-object interaction detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 7
- [25] Gyeongsik Moon, Heeseung Kwon, Kyoung Mu Lee, and Minsu Cho. Integralaction: Pose-driven feature integration for robust human action recognition in videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. 1

- [26] Boxiao Pan, Haoye Cai, De-An Huang, Kuan-Hui Lee, Adrien Gaidon, Ehsan Adeli, and Juan Carlos Niebles. Spatio-temporal graph for video captioning with knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020. 1
- [27] 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*, 2023. 3, 7
- [28] Xian Qu, Changxing Ding, Xingao Li, Xubin Zhong, and Dacheng Tao. Distillation using oracle queries for transformer-based human-object interaction detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 3
- [29] 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*, 2021. 3
- [30] T-YLPG Ross and GKHP Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2017. 5
- [31] Zehua Sun, Qihong Ke, Hossein Rahmani, Mohammed Bennamoun, Gang Wang, and Jun Liu. Human action recognition from various data modalities: A review. *IEEE transactions on pattern analysis and machine intelligence*, 45(3): 3200–3225, 2022. 1
- [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*, 2021. 7
- [33] Francesco Tonini, Nicola Dall’Asen, Cigdem Beyan, and Elisa Ricci. Object-aware gaze target detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. 5, 6
- [34] Danyang Tu, Wei Sun, Guangtao Zhai, and Wei Shen. Agglomerative transformer for human-object interaction detection. In *Proceedings of the IEEE/CVF international Conference on Computer Vision*, 2023. 3, 7
- [35] Oytun Ulutan, ASM Iftekhar, and Bangalore S Manjunath. Vsgnet: Spatial attention network for detecting human object interactions using graph convolutions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020. 3
- [36] Guangzhi Wang, Yangyang Guo, Ziwei Xu, and Mohan Kankanhalli. Bilateral adaptation for human-object interaction detection with occlusion-robustness. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 3
- [37] 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*, 2019. 1, 3
- [38] Eastman ZY Wu, Yali Li, Yuan Wang, and Shengjin Wang. Exploring pose-aware human-object interaction via hybrid learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 7
- [39] Chi Xie, Fangao Zeng, Yue Hu, Shuang Liang, and Yichen Wei. Category query learning for human-object interaction classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 7
- [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. 7
- [41] Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. Exploring visual relationship for image captioning. In *Proceedings of the European Conference on Computer Vision*, 2018. 1
- [42] 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. In *Advances in Neural Information Processing Systems*, 2022. 7
- [43] 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*, 2023. 7
- [44] 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. In *Advances in Neural Information Processing Systems*, 2021. 7
- [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*, 2022. 3, 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*, 2023. 3, 7
- [47] Shanshan Zhang, Rodrigo Benenson, Mohamed Omran, Jan Hosang, and Bernt Schiele. Towards reaching human performance in pedestrian detection. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):973–986, 2017. 1
- [48] Shanshan Zhang, Jian Yang, and Bernt Schiele. Occluded pedestrian detection through guided attention in cnns. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018. 1
- [49] Yubo Zhang, Pavel Tokmakov, Martial Hebert, and Cordelia Schmid. A structured model for action detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019. 1
- [50] 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*, 2022. 3, 7

- [51] Xubin Zhong, Changxing Ding, Xian Qu, and Dacheng Tao. Polysemy deciphering network for human-object interaction detection. In *Proceedings of the European Conference on Computer Vision*, 2020. [1](#), [3](#)
- [52] 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*, 2021. [3](#)
- [53] Penghao Zhou and Mingmin Chi. Relation parsing neural network for human-object interaction detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019. [1](#), [3](#)
- [54] 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*, 2021. [6](#), [7](#)